ELEVENTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION

Held under the patronage of the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT)

MAY 7– 12, 2018

PHOENIX SEAGAIA CONFERENCE CENTRE
Miyazaki**, JAPAN**

CONFERENCE ABSTRACTS

Editors: Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis, Takenobu Tokunaga

Assistant Editors: Sara Goggi, Hélène Mazo

LREC 2018, ELEVENTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION

Title: LREC 2018 Conference Abstracts

Distributed by:
ELRA – European Language Resources Association
9, rue des Cordelières
75013 Paris
France

Tel.: +33 1 43 13 33 33
Fax: +33 1 43 13 33 30

www.elra.info and www.elda.org
Email: info@elda.org and lrec@elda.org

ISBN 979-10-95546-00-9
EAN 9791095546009

Introduction to LREC 2018 by Nicoletta Calzolari
Chair of the 11th edition of LREC
ELRA Honorary President

Welcome to the 11th edition of LREC in Miyazaki, first LREC in Asia!

LREC 20th Anniversary

It is the LREC 20th Anniversary and LREC has become one of the most successful conferences of the field. Data are pervasive in Natural Language Processing and Language Technology: we call our data Language Resources (LR). But when LREC was started by ELRA, in 1998 in Granada, from an idea of Antonio Zampolli and Joseph Mariani, it was really a new adventure and a challenge. There were well established big conferences but he thought that the new emerging field of Language Resources deserved its own dedicated forum. In the keynote talk I gave at LREC1998 I could say: "the infrastructural role of Language Resources as the necessary common platform on which new technologies and applications can be based is nowadays widely recognised." This could not have been said only few years before. I had the pleasure and the honour of being involved in LREC from the beginning, first as member of the Program Committee and since 2004 as Conference Chair.

LREC is probably the most influential ELRA achievement, and a service with the major impact on our community. Also through LREC, ELRA contributes to shape our field, making the Language Resource field a scientific field in its own right.

Why LREC in Asia this time? AFNLP (the Asian Federation of NLP) asked us if we could hold an LREC in Asia as the best instrument to promote Language Resources in Asia. We were glad to accept this challenge and here we are.

Some LREC2018 figures

As expected given the change in continent, we did not break any record this time, but the figures are not far from the previous. We received 1102 submissions for the main conference, 34 workshop proposals and 8 tutorial proposals.

A very large part of our community was involved in the reviewing effort, to be able to assign few papers per reviewer: 1263 colleagues accepted to act as reviewers (more than in 2016) out of 1796 invited (268 declined and 265 unfortunately not answering). Few reviewers did not complete the task (only 26 reviews missing, not so bad), but knowing that this always happens we recruited some pinch-reviewers able to act at the last moment: a good move to keep for the future.

The Program Committee has also been enlarged with 3 colleagues from Japan and one from USA. We had as usual a very hard job, examining about 3300 reviews, to understand – beyond the scores and in particular when they greatly differed – the relevance, the novelty, but also the appropriateness for an oral or poster presentation. I am sure we made mistakes, every reviewing effort is not immune from subjectivity, but as usual we discussed in a face to face meeting not only general policies, criteria and how to be consistent, but also borderline cases to arrive at agreed decisions. Overall we all believe we received in average good submissions. We have in the main program 718 papers: 188 Orals and 530 Posters.

We also have 29 Workshops and 5 Tutorials.

I am proud that around 1100 participants have already registered at the end of April, similar to last time. They come from 63 countries. The Japanese are the largest group and in general there is a larger participation from Asian countries, in particular China, as we obviously hoped.

These figures have a clear significance. The field of Language Resources and Evaluation is very alive and constantly flourishing.

LREC acceptance rate: a motivated choice for an inclusive conference

The LREC acceptance rate, 65% this year, is different from other major conferences but for us it is a motivated decision. This is one of the reasons why LREC succeeds to provide a comprehensive picture of the field and to show how it is evolving. For us it is important not only to hear about new methodologies but also to understand how various methods or resources are able to spread, for which purposes, usages, applications, and for which languages. Multilingualism – and equal treatment of all languages – is an essential feature of LREC, as it is the attempt of putting the text, speech and multimodal communities together as well as academics and industrials. LREC wants to be an "inclusive" conference.

Quality is not undermined by our acceptance rate: in 2017 Google Scholar Metrics h5-index, LREC ranks 4th in Computational Linguistics top conferences (5th considering ArXiv which is the first).

LREC2018 Novelties

Industry Track

Because of the interest in joining forces between academy and industry, this time we decided to experiment with a new Industry Track. We spoke about this at last LREC with Linne Ha from Google and we asked her if she wanted to organise it for LREC2018.

Special Speech Session

A special session on "Speech resources collection in real-world situations" was proposed to us by Kikuo Maekawa and Yuichi Ishimoto (National Institute for Japanese Language and Linguistics): we gladly accepted also to strengthen the participation of the speech community at LREC.

Oriental-COCOSDA Conference

Also O-COCOSDA is organised together with LREC. We spoke with Satoshi Nakamura, its chair, at last LREC and he kindly offered to organise it jointly with LREC. We are very pleased of this also because it is another opportunity to reach the Asian speech community.

ELRA Individual Members Assembly

ELRA has recently introduced "individual membership" in addition to institutional membership. This was decided to give a voice inside ELRA to the large LREC community and offer them its services. The first assembly of ELRA individual members is held on the first day of the conference.

The LREC Club

From the answers received, it seems that the LREC Club of those who attended all editions, the really faithful ones, is composed of 23 members. I want to thank them for their loyalty!

LREC2018 Trends

I quickly sketch here, as I always do, my perception – subjective and impressionistic – of LREC2018 trends and how certain topics fluctuate from an LREC to the other. The comparison with previous years shows the topics with steady progress, or even great leaps forward, the stable ones and those more affected by the fashion of the moment.

Trends in LREC2018 topics

Among the areas that continue to be trendy and are even increasing I can mention:
- Less-Resourced Languages
- Social Media analysis, appearing in 2012 and since then constantly growing
- Semantics in general and in particular Sentiment, Emotion and Subjectivity
- Information extraction, Knowledge discovery, Text mining are booming
- Lexicons (in its various forms)
- Discourse, Dialogue, Conversational systems and Interactivity
- Multimodality, also for Less-Resourced languages
- Tools, Systems, Applications for various purposes: Question Answering, Summarisation, etc.
- Evaluation methodologies
- Computer Aided Language Learning

Stable "usual" topics, some very well-represented, others in the medium/low range, are:
- Infrastructural issues, policies, strategies and Large projects: topics that receive special attention at LREC, differently from other major conferences
- Corpus creation, annotation, use, …
- Speech related topics, a little increasing but not as much as we would like
- Sign language (also a very successful workshop)
- Crowdsourcing
- Anaphora and Coreference
- Temporal and Spatial annotation

New trends for this LREC:
- Digital Humanities (new for LREC in 2016, now increased)
- Bibliometrics, Scientometrics, Infometrics
- Language Modelling

Decreasing topics with respect to the past, even if some still numerous:
- Grammar and syntax and also Treebanks that had a big increase in 2016
- Multilinguality and Machine Translation, very high in 2016
- Ontologies
- Standards and metadata are much less represented
- Linked data, a new topic in 2014, seems no longer so fashionable
- Web services and workflows also no longer so popular

The recognition given by the LR community to infrastructural issues, strategies and policies may be also due to the fact that we must often work in large groups, for many languages, we must build on each other work, connect various resources and tools, make available what already exists and use standardised formats. Infrastructures (on many dimensions) are really needed for our field to progress: to pay proper attention to these issues is another distinguishing feature of LREC.

LREC-related initiatives

Proceedings in Thomson Citation Index

Since 2010 the LREC Proceedings are included in CPCI (Thomson Reuters Conference Proceedings Citation Index): an important achievement, providing a better recognition to all LREC authors and useful in particular for young colleagues.

LRE Journal and LREC

After each LREC we ask to the authors of papers suggested by the 3 reviewers as appropriate for LRE if they want to submit an extended version to the *LRE journal*, coedited by Nancy Ide and myself. I am glad to report that also the journal has a large and increasing number of submissions, testifying the great interest for the field of LRs and Evaluation.

Citation of Language Resources

Also this year we encouraged citations of LRs in a special References section (introduced in 2016), providing recommendations on how to cite. I hope this becomes normal practice, to keep track of the relevance of LRs but also to provide due recognition to those working on LRs.

LRE Map and Share your Language Resources

As usual we encouraged descriptions of LRs in the LRE Map, an innovative instrument introduced at LREC2010 with the aim of monitoring the wealth of data and technologies developed and used in our field. And we ask, since 2014, to share the LRs with all the community.

In this LREC about 1000 LRs have been described in the Map. Just few hints at some data in the 2018 Map: WordNets, Wikipedia, Prague TreeBank are the most cited LRs; Corpora are by far the most frequent type (half of the LRs); and about 85% of the LRs are in some way available (not bad).

Replicability of research results

I believe that research is strongly affected also by infrastructural (meta-research) activities as those mentioned above. With these initiatives I hope we are able to promote in our field what is in use in more mature disciplines, i.e. ensure proper documentation and reproducibility of research results as a normal practice. ELRA and LREC are thus influential in strengthening the Language Resources and Evaluation scientific ecosystem and fostering sustainability.

Acknowledgments

As usual, it is my pleasure to express here my deepest gratitude to all those who made this LREC2018 possible and hopefully successful.

I first thank the Programme Committee members, not only for their dedication in the huge task of selecting the papers, the workshops and tutorials, but also for the constant involvement in the various aspects around LREC. I wish to thank each of them, the new ones: Chris Cieri, Koiti Hasida, Hitoshi Isahara, Take Tokunaga, and obviously the “old” ones: Khalid Choukri, Thierry Declerck, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis. A particular thanks goes to Jan Odijk, who has been so helpful in the preparation of the program.

I thank ELRA and the ELRA Board: LREC is a major service from ELRA to all the community!

A very special thanks goes to Sara Goggi and Hélène Mazo, chairs of the Editorial Committee, for the dedication and competence in managing the so many tasks they have in their hands, and the capability to tackle the many big and small problems of such a large conference (not an easy task). They are the two pillars of LREC, without whose commitment for many months LREC

would not happen. So much of LREC organisation is on their shoulders, and this is visible to all participants.

I am especially grateful to Hitoshi Isahara and Kyoko Kanzaki, the Local Committee, for their great efforts in dealing with so many local matters and for their patience and true commitment, looking at so many details. We owe them a lot in organising a successful LREC.

My appreciation goes also to the distinguished members of the Advisory Board, chaired by Makoto Nagao, for their support and precious advices.

I am very grateful to the Local Liaison Committee, representing a number of Asian associations and organisations that have supported LREC, in particular with dissemination tasks.

I express my great gratitude to the Sponsorship Committee, and to all the Sponsors that have helped with financial support, believing in the importance of our conference.

I thank the Japanese Ministry of Education, Culture, Sports, Science and Technology for its precious support.

I am particularly grateful to the local authorities, the Governor of Miyazaki prefecture and the Mayor of Miyazaki city, very supportive of LREC since my first visit in 2016. We met them, and other local authorities, several times. We thank them also for their financial support.

In my many visits to Miyazaki I have been impressed by the great sense of hospitality of locals, and among them I wish to thank at least Manmatsu Hayashi and Rie Saita for their great help and kindness.

Also on behalf of the Program Committee, I praise our impressively large Scientific Committee. They did a wonderful job.

I thank the workshop and tutorial organisers, who complement LREC of so many interesting events.

I thank the organisers of the Industry Track, of the Special Speech session and of O-COCOSDA.

A big thanks goes to all the LREC authors, who provide the “substance” to LREC, and give us such a broad picture of the field.

This time I really want to thank also Softconf (Rich Gerber and Paolo Gai) and their constant efforts to make START a better tool for us. I greatly appreciated the new feature for plagiarism detection: I must say it was useful to detect some paper too similar to others …

I thank the European Commission for the interest in our conference, and hope that funding agencies will be impressed by the quality and quantity of initiatives in our sector that LREC displays, and by the fact that the field attracts the best groups of R&D from all continents. The success of LREC for us means the success of the field of Language Resources and Evaluation.

I finally thank the two institutions that always dedicate a great effort to LREC: ELDA in Paris and ILC-CNR in Pisa. Without their commitment LREC would not be possible. The last, but not least, thanks are thus, in addition to Hélène Mazo and Sara Goggi, to all the others who – with different roles – have helped and will help during the conference: Roberto Bartolini, Damien Bihel, Irene De Felice, Riccardo Del Gratta, Pawel Kamocki, Valérie Mapelli, Monica Monachini, Vincenzo Parrinelli, Vladimir Popescu, Valeria Quochi, Caroline Rannaud, Alexandre Sicard.

And lastly, my final words are for all the LREC2018 participants, the true protagonist of LREC. Now LREC is in your hands. I hope that you discover new paths, that you perceive the vitality and strength of the field, that you have fruitful conversations (conferences are useful also for this) and most of all that you profit of so many contacts to organise new exciting work and projects in the field of Language Resources and Evaluation … which you will show at the next LREC.

This Japanese LREC in Miyazaki has a sort of Mediterranean flavour, typical of LREC. I am sure you will appreciate the Japanese great hospitality and kindness. And I hope that Miyazaki will enjoy the invasion of LRECers!

With all the Programme Committee, I welcome you at LREC2018 and wish you a very fruitful Conference.

Enjoy LREC2018 in Miyazaki!

Message to LREC 2018 participants by Henk van den Heuvel ELRA President

Dear Colleagues and Friends,

Twenty years of LREC! It is my honour and pleasure to welcome you to this 11th edition of our successful Conference. Welcome to Miyazaki!

We are also very grateful with our guests representing the European Commission. Your presence here is deeply appreciated. Especially we welcome, Gael Kent, Director Data at the European Commission- DG CONNECT in Luxemburg. We are looking forward to your speech.

We are very honoured to have literally in our midst Prof. Makoto Nagao, Professor Emeritus of Kyoto University. We greatly admire your contributions to such various fields as Machine Translation, Natural Language Processing, Pattern Recognition, Image Processing and Library Science.

After 20 years LREC we have broken with the tradition to convene around the Mediterranean area, and look for another venue to meet and network. Honestly, we see this as an exceptional move motivated by our deep desire to intensify the ties with our Asian colleagues as we know them from the Asian Federation of Natural Language Processing (AFNLP), the Board of which is also closely involved in the organisation of this LREC through the Local Liaison Committee. We are very pleased to see so many of our Asian colleagues here in Miyazaki.

As President of ELRA it is my duty and pleasure to point out a couple of developments that are taking place in our Association. Already in 2012 one of my predecessors, Stelios Piperidis, referred to the dazzling speed of changes in which our community is finding itself. In his opening speech at LREC 2012 he also mentioned the upcoming of data-driven techniques and numerical and learning methods. In our days we see how algorithms and techniques developed in the area of Artificial Intelligence have come to play a paramount role in the area of Language and Speech technology. This technology puts special demands on the amount and preprocessing of Language Resources for training and testing purposes. Large amounts of data are collected from the web and continuously processed and used for application refinement. Now, in this rapidly changing field, ELRA has to find its way as one of the traditional sustainable key-players in language resources management and intermediary between stakeholders. It is evident that LRs remain essential also in our time, it is also evident that well-targeted annotated resources remain essential for supervised training approaches. Therefore, there remains an important role for ELRA as a sustainable LR broker offering relevant and high quality resources both to academia and commercial parties.

However, the changes that we see around us force us to continuously reflect on our *raison d'être* for our members in consideration of what their demands are for LRs and in terms of the services we offer around them. As a result of that, ELRA's Board has introduced important changes in its membership policy.

First of all, to stimulate continuity for our institutional members we have introduced a discount on membership fees upon membership continuation, starting with a discount of 15% for the second year up to 30% for the third year and following. Second, we have equalized the fees for EU and non-EU members to the EU-members fee. Last but not least, as of January 2018 ELRA

has introduced individual membership. An event such as LREC shows how vivid and productive the community around LRs is, and advocates for establishing a permanent link within this community, not only a biennial meeting point. For this reason ELRA has decided to open up its memberships for individuals, too, and to offer this membership with special services and benefits, of which the reduced registration fee is the one now most salient.

Employees of institutional ELRA members are also individual ELRA members if and when they want to use ELRA member services (including discount on LREC registration fees). They will not have to pay the individual membership fees as well since their organization covers for that.

In addition, one position in the ELRA Board will be reserved for a representative from the individual members, and this member is elected by the individual members only. This Board member has the same rights as the other ELRA Board members on all issues related to Board matters.

There will be a General Meeting for individual members at each LREC where they can convene with their representative and the Board to discuss ELRA matters concerning individual members. This meeting will be organized for the first time in this LREC 2018, namely this very afternoon at 18:00. The content of the meeting is an interesting mixture of relevant issues from the ELRA board, an inventory of wishes from individual members, and a self- introduction of Board applicants.

You are all invited to attend this first ELRA membership meeting, where we will tell more about the new membership policy, the special services for members and the election of the new Board member. We have sent out an invitation and an agenda for this.

Another observation that requires our persistent attention is that there are many players offering LRs both at the national and international level, and this landscape is becoming quite diffuse. This implies that we need to identify and re-identify times and again what our, ELRA's, position is compared to other LR brokers. It is ELRA's firm belief that this can best be done through cooperation. In this way we have set up a successful cooperation with for instance, LDC, by identifying the differences in membership policies, LR production and distribution strategies, and using each other's strengths in cooperation. In the same spirit ELRA has now set up a Collaboration Agreement with CLARIN ERIC. In this Collaboration Agreement we have clearly identified where our mutual and complementary strengths are and how we can bring these together to the benefit of both organisations. The objective of such agreements is not that one organization becomes part of the other but that both remain independent whilst joining forces. Indeed, here we see an important role for our association in facilitating synergies.

Another example of such a synergy has been established in our Special Interest Group for Under-resourced Languages, SIGUL. Created in April 2017, SIGUL is a joint Special Interest Group of the European Language Resources Association (ELRA) and of the International Speech Communication Association (ISCA). Through its establishment of the Special Interest Group on Under-resourced Languages, ELRA reasserts its active involvement in contributing to enhance the support for the languages with little or no technological support.

I would like to take the opportunity to thank all those who have worked so hard to make this conference a fantastic event: the LREC Programme Committee, chaired by Nicoletta Calzolari, the Scientific Committee, the Conference Editorial Committee headed by our LREC cornerstones Sara Goggi and Hélène Mazo, the International Advisory Committee chaired by Prof. Makoto Nagao, the group in Pisa, Khalid Choukri and the ELDA staff in Paris, the Local Committee headed by Prof. Hitoshi Isahara and Dr Kyoko Kanzaki. Each one of them in his/her own role has been taking care of the incredible amount of issues that emerge when undertaking

the organisation of such a complex and demanding conference as LREC. Our particular thanks go to our sponsors and supporters.

We thank workshop and tutorial organizers, project consortia participating in the HLT Village; you have all exceeded yourselves once more to make this LREC such a great event.

Dear LREC Participants, in the end this is your conference. With your active participation in the oral sessions, your lively discussions with the presenters at the poster sessions, your visits to the HLT Village and Exhibition Boots and participation in the Industry Track I am confident you will make LREC 2018 yet another success.

Introductory message of Khalid Choukri,
ELRA Secretary General
ELDA Chief Executive Officer

ELRA and ELDA are very pleased to welcome you in Miyazaki to this 11th LREC to celebrate the 20th anniversary of LREC with all of you this week.

On behalf of the ELRA/ELDA team I would like to share with you some news on the activities we conducted since the last LREC in Portorož (Slovenia).

The Declaration of Granada

But first let me to share some feelings about this special LREC with you, as we are celebrating the 20th anniversary of this major forum established in 1998 in Granada (Spain), organized for its 11th edition, here in Japan.

Soon after the establishment of ELRA in 1995, its Board realised that, at that time, the language resources and the evaluation of language technologies were given very little attention at the main events. Today, we are glad that such message is spread widely and is endorsed by the major conferences in which special sessions are expressly devoted to Language Resources and Evaluation!!

Remember the first LREC, remember Granada, not only the Alhambra! With over 400 participants instead of the expected 100 attendees, we realized the importance of such forum for the community. This was confirmed over the years by a steady attendance of 1200 participants to the last editions of LREC.

I would like to take this opportunity to go back to the spirit of Granada, paying a tribute to those who were behind it, Professors Antonio Zampolli and Angel Martin Municio. I would like to bring up one of the major outcomes of that first event: “the declaration of Granada”. Its recommendations are still relevant and topical, more urgent than ever to implement.

The declaration of Granada[1] comprised 10 articles. I am highlighting and commenting here some of the crucial ones that we can continue to endorse today:

- "**At this moment, language resources are one indispensable key to unlock the potential of the global information Society**"

We are still facing this issue 20 years later and if we agree that the Information Society has made tremendous progress with the emergence of social networks which have strengthened links within and between communities, social or commercial activities cross borders are still hindered by language barriers. In 2015, surveys mentioned that 24 languages are used in LinkedIn user interfaces, 48 on Twitter, 91 on Google Translate (as pairs for its translation of content and now about 103), over 150 on Facebook, just over 300 in Wikipedia. These numbers may seem impressive, but remember that this is **out of 7097 living languages or 3,909 with writing systems.** And most of these languages are used in interfaces with automatic processing of content used in Search and/or MT only. Language Resources are essential assets. Back in

[1] Granada Declaration: http://www.elra.info/media/filer_public/2013/09/06/v3n3.pdf)

Granada, we stated that "They constitute an essential infrastructure". Such infrastructure is missing for a huge number of languages. The LRE Map service provided by ELRA, inventorying the LRs reported in major conferences, continue to expose the existing gaps.

- **"All sectors of society, and all languages, have an interest in seeing these resources developed, for a variety of purposes, economic, social, industrial and cultural."**

ELRA continues to promote the concept of Basic Language Resource Kit, a Kit that would help process every language for (at least) the basic NLP functions. We stressed the importance of this approach to policy makers, emphasised the need to support small communities, and mentioned the lack of interest from private sector for non-lucrative/non-strategic languages. We also insisted that such "*core language resources should remain in the public domain*" to ensure a wide use by both research and development stakeholders. Reviewing the current situation at major data centers and repositories, we can barely count more than 100 different languages, often with scarce resources (many speech resources for the major languages, very few treebanks, very few aligned corpora, mostly aligned with English, etc.)

- **"For each language, there is a need for strategy to co-ordinate existing resources and create new ones."**

ELRA, along with LDC, their partner in the USA, did their best to offer distribution/sharing channels for Language Resources produced within publicly funded projects and some offered by private bodies. However the identified resources represent less than 15% of what exists. Coordination of the distribution but also documentation and production, have proved to be challenging. We still feel it is crucial to coordinate building roadmaps for every language and enhance the involvement of local public and private bodies. It is also essential to continue international cooperation to disseminate the know-how acquired for a given language. We are glad that a conference like LREC contributes to sharing such expertise and value the implication of governmental (regional and national) and international bodies.

We introduced the International Standard Language Resource Number (now part of the activities of the International Standardisation Organisation, ISO TC37/SC4) to assign a unique identifier with each identified Language Resource to improve the way we reference it (this is also part of the LREC submission process that distinguishes Bibliographical data from LR data). The idea is not only to provide an ID, unique and persistent, wherever the LR is stored, even for those LRs on local servers outside the Internet. This is an uphill struggle but we are convinced that it is an important step in our work to improve the identification of existing resources, the assessment of LR impact factor as well as the citation mechanism.

- **"When resources have been created, there is a continuing requirement for support and maintenance."**

This is a key part of our mission and we tried to convince data producers and funders to account for the necessary maintenance of and support for Language Resources. We introduced the validation process and the "bug" reporting mechanism, as part of ELRA procedures, to encourage sharing experiences on the use of LRs and their enhancement over time. We still face funding scenarios that provide subsidies for data production and not for other issues like IPR clearance, documentation, sharing, maintaining, etc. In Granada, we anticipated that resources would undergo some repurposing with the new uses that emerge and we insisted on the need to envisage a wide range of applications on the basis of the same resources. The community seems to be sensitive to this, but some legislators are debating the adoption of more legal constraints. We need to join forces to convince funders and decision makers about the importance of more openness and long term policies. The introduction of the Data Management

Plan (DMP) by ELRA, and soon the DMP Wizard, will help each data manager to adopt up-to-date standards and best practices for data management.

- **"Understanding of the role, usefulness and optimum means of preparation for language resources is a research theme in itself."**

Over the last decades, and especially within the last 3-4 years, we have seen an impressive breakthrough in the HLT field. The new data-intensive machine learning and the computing capabilities, are proving the crucial usefulness of LRs. Making LRs widely available is the core mission of a few organisations. ELRA is very happy to be among these organizations and is making the necessary investments to acquire more expertise to cost-effectively produce and share LRs. The setup of an internal legal team is helping to shed light on a large number of legal issues that impede the use/re-use of LRs. Working on standards is also an important aspect to help facilitate the interoperability and sharing of data. One of our mottos was that "Common evaluation requires common standards". We still feel that common tasks in the "challenges" and evaluation campaigns are essential instruments to assess progress, share knowledge, and improve cooperation. It is a pity that many "Evaluation campaigns" are happening with very little coordination which makes them hard to find for new comers.

Granada was 20 years ago and we see that some visionary recommendations are still needed today. A multilateral, concrete, and lasting cooperation remains on top of our action.

ELRA activities since 2016

Now allow me to get back to ELRA activities carried out over the last couple of years.

We continue our actions on data sharing, through the identification, negotiation, and distribution agreements with right holders when necessary. We continue to produce resources for projects as well as for partners. Our policy remains consistent: whenever the data is offered to the community, after the shortest possible embargo period, the costs for partners are set to production costs. This position remains fundamental to our policy. We continue to invest in research and development of tools to improve and automate our production procedures. Most of our tools are shared as open source packages.

We continue also to work on our quality control methodologies so as to supply validated resources with validation procedures that guarantee the adequacy of the produced datasets with respect to the initial specifications and the state of the art.

To ensure an efficient distribution of Language Resources, ELRA has migrated its catalogue of resources to a new platform, based on e-Commerce features, redesigned with a new interface and an improved navigation. This foreshadows further developments that will incorporate e-licensing, e-payment and e-delivery of resources.

ELRA continues to support the set-up of LR repositories for data deposit by third parties. Based on its involvement in the jointly-developed META-SHARE platform, we continue the promotion of such efforts to ensure that the major data holders adhere to some common practices. A new repository was set up as part of an EU service contract to store data for MT provided by the public sector. Such initiative is now spreading across Europe, and a coordination action is establishing local repositories (known as Local Relay Stations). If we succeed to set up such stations for each country in order to collect all language datasets produced by translations services and secure these for MT training and tuning, one can anticipate good progress for these languages and domains. The repositories can accommodate any Language Resource modality.

If the establishment of such a local repository is of interest to your organization and your network, let us discuss how to work on it together.

As part of this process, we continue to work on all issues related to sustainability and preservation of data for the generations to come.

An updated ELRA Data Management Plan is made available and reviews all necessary aspects for an optimal management of resources with an easy-to-use checklist. We are working to automate the customisation of such DMP for each project. Our members will benefit from this automatic DMP Wizard, accompanied with the support of our experts, free of charge. We hope that such approach will improve sustainability and preservation of Language Resources but also make them easy to identify.

ELRA continues to be involved in the new trends in HLTs. It continues to support the new trends in MT. Many of our projects (some of which are funded under a European Program known as Connecting Europe Facility (CEF) focus on data production, including via requests for donations from translation services, but also crawling of adequate data to which we have access and re-use rights. Many resources come from organizations that belong to the Public Sector. A directive (called Public Sector Information directive, PSI) entered into application in the European Union, similar rules exist in many other countries, stating that publicly produced data should be made publicly available. This makes some of the resources needed by our community (e.g. textual corpora) available for new domains and new genres. Some geographical areas offer a multilingual environment (EU, India? South Africa, etc.), and hence more resources should be available for MT development.

Unfortunately there are still important legal restrictions on the re-use of data, even for research purposes. We continue to vilify the current legal framework, in particular in Europe, e.g. the European Union is working on a new directive on copyright in the Digital Single Market. The initial proposal for this act contained a mandatory exception for text and data mining carried out by research institutions. However, the current debates within the European decision makers seem to suggest that the exception will fall short of meeting the objective of the exception. The beneficiaries of the new exception may be limited to public research institutions, and – more importantly – 'lawful access' will be a prerequisite for data mining, which will probably result in wider implementation of digital protection measures by right holders. It is unlikely to get the exception for research that we claim since years now as a fair use doctrine for research purposes (that remains the privilege of a few countries).

The current legal framework has a strong impact on the capacity of the community to produce IPR cleared and sharable data. ELRA heavily invested in legal training and has been, for many years now, one of the few organizations that works both with in-house legal experts and a network of external practitioners/lawyers.

Another critical novelty in Europe is the new legal framework governing the processing of personal data. It goes beyond the users expectations, for more ethical behaviour on the management of their data. This may hinder the new developments of resources and technologies (e.g. Crowdsourcing activities). The new regulation (General Data Protection Regulation (GDPR)) will impose more restrictions on managing several aspects of data e.g. data protection by design and by default, privacy impact assessment, pseudonymisation and anonymization, before the data can be shared (this will of course impact also production, repackaging, repurposing of data).

To share information on these matters, a dedicated workshop on legal and ethical issues continues to be organized within LREC and will be held this week as well.

Of course, ELRA does not focus on EU issues and EU languages only (we distribute resources for more than 70 languages). In 2017, ELRA entered into an important agreement with the International Speech Communication Association (ISCA[2]) to join forces in the promotion of activities related to the Less-Resourced Languages (LRL). ELRA and ISCA agreed to merge their groups and set up a join Special Interest Group for Under-resourced Languages (SIGUL[3]). Co-chaired by a representative of ELRA and a representative of ISCA, SIGUL will continue to organize events for the LRL and encourage cooperation actions to support these languages.

As you may know, United Nation General Assembly proclaimed 2019 as the International Year of indigenous Languages. UNESCO is leading the corresponding events. ELRA proposed to organize an important international event related to HLT and Indigenous languages. We hope to draw attention to the importance of HLT and LRs for the preservation and development of local cultures and put under spotlights the role our community could play for these languages.

We continue to develop the LRE Map application. LRE Map was established to reference all LRs described by authors when submitting papers to conferences and journals. Started with LREC, it is used by other events but not as widely as we hope. In addition to identifying over 7000 instances of LRs, it helps identify existing gaps for languages lacking such modalities and ensure a minimal cooperation when planning new productions. If you are involved in the organisation of a conference, let us see how we can work together.

ELRA is also taking part in several standardisation activities. It is naturally involved in ISO/TC37/SC 4 on Language Resource management but also on ISO/IEC JTC1/SC35 about user interfaces and accessibility. ELRA brings its knowledge of the HLT field to ensure that all ICT services and products are accessible to all, in particular to users with specific needs. Some of the HLT applications are offering valuable services when converting speech into text, text into speech, sub-titling/captioning audio-visual streams, providing audio descriptions, translations (e.g. subtitles), easy-reading features (both in mono- and multilingual contexts). Such services are valuable to everyone and not only hearing or visually impaired users. Translation from text or speech to Sign languages is a big challenge that many partners are working on and ELRA will support them.

As a conclusion to my message, I would like to reiterate my statement uttered at almost all LRECs since 1998. Please remember that we can help you share your data for all types of use. We can work out a contractual framework that suits your expectations, including adopting very permissive licences and a free-of-charge policy. We can guarantee the availability as well as the sustainability of your resources. During the conference, an ELRA booth is available where we will be happy to interact with you on such topics.

About 10 years ago, we identified about 20 resources, some were on the web, others well known to the community. We keep monitoring their availability. Believe it or not, about 30% disappeared and these are not necessarily the ones that were obsolete and useless. Some right holders also disappeared and the “orphan” resources with them.

Acknowledgments

Finally, I would like to express my deep thanks to our partners and supporters, who throughout the years make LREC so successful. I would like to thank our Sponsors: Google, Amazon, Arcadia, EML (European Media Laboratory GmbH), GSK Language Resources, Riken-AIP, Yahoo Research Japan, and the publisher Hituzi Syobo our media sponsor: MultiLingual Computing, Inc.

[2] https://www.isca-speech.org/iscaweb/index.php/about-isca
[3] http://www.elra.info/en/sig/sigul/

I also would like to thank the HLT Village participants, we hope that such gathering offers the projects an opportunity to foster their dissemination and hopefully to discuss exploitation plans with the participants.

I would like to thank the Local Advisory Committee. Its composition of the most distinguished personalities of Japan denotes the importance of language and language technologies for the country.

I would like to thank the LREC Local Committee, chaired by Prof. Hitoshi Isahara and the LREC Local Organizing Committee, for providing support to the organization of this LREC Edition in Japan.

Finally I would like to warmly thank the joint team of the two institutions that devoted so much effort over months and often behind curtains to make this one week memorable: ILC-CNR in Pisa and my own team, ELDA, in Paris. These are the two LREC coordinators and pillars: Sara Goggi and Hélène Mazo, and the team: Roberto Bartolini, Damien Bihel, Irene De Felice, Valérie Mapelli, Monica Monachini, Vincenzo Parrinelli, Vladimir Popescu, Caroline Rannaud, and Alexandre Sicard.

Now LREC 2018 is yours: we hope that each of you will achieve valuable results and accomplishments. We, ELRA and ILC-CNR staff, are at your disposal to help you get the best out of it.

Once again, welcome to Miyazaki and Japan, welcome to LREC 2018

Message to LREC 2018 participants by of Hitoshi Isahara and Kyoko Kanzaki, Chairs of the Local Committee

Welcome to Japan! Welcome to Miyazaki!!

On behalf of the whole local team, we would like to extend our warmest welcome to all of you participating in this 11th edition of LREC in Miyazaki, Japan!

You may know personally some Japanese researchers in the fields which are involved in LREC, and also may know that Japan is a country with four distinctive seasons, blessed with beautiful nature, world heritage sites, rich culture and respected traditions. However, we are sure that most of you are not familiar with Miyazaki.

Miyazaki has beautiful coastline facing the Pacific Ocean, and has many shrines involved with the myth of the birth of Japan, such as Miyazaki Shrine sacred to Emperor Jinmu, supposedly the first Emperor of Japan. Most stories in the mythology associated with the creation of Japan and the origin of the imperial line took place in Miyazaki Prefecture on the island of Kyushu.

You can visit scenic places and historical places in Miyazaki in this occasion, or you will be able to visit here again with your friends and family.

During your stay here, we would like you to experience Omatsuri, festival in Japan. You can enjoy Shrine Maiden Dance, Kagura performance (a sacred music and dancing performance dedicated to the Shinto gods) and local cuisine during reception. Lunch will be served in stall style.

On this occasion, we would like to thank all the supporters in Japan from the bottom of **our** hearts.

We first appreciate the patronage of the Ministry of Education, Culture, Sports, Science and Technology (MEXT) towards LREC2018.

We thank Miyazaki Prefecture and Miyazaki city for their continuous support from the beginning of LREC's venue selection process. We offer the Governor, Mr. Shunji Kono, and the Mayor, Mr. Tadashi Tojiki our heartfelt thanks for their great kindness and efforts. Thanks to their generous understandings, preparation of LREC went smoothly.

We also thank Miyazaki Convention and Visitors Bureau for its financial and human support. We are grateful to Mr. Mitsunori Mera, Mr. Toshiaki Tomitaka and Mr. Manmatsu Hayashi.

We are sure that all participants satisfy high quality service of the conference venue, Phoenix Seagaia Resort. We appreciate very hard work by Mr. Hirofumi Matsunaga, Mr. Manmatsu Hayashi, Ms. Rie Saita and Mr. Satoru Kamibayashi to meet the requirements which LREC indicated.

One of the highlights of LREC2018 is the Reception at Miyazaki Shrine which is sacred to the first Emperor of Japan. We are grateful to its Chief Priest, Mr. Hidekiyo Sugita, for accepting us.

Participants can enjoy local liquor at the Reception and Gala Dinner. Let's thank to Miyazaki Sake Brewers Association.

We would like to thank Ms. Rika Kubota for organizing interpreter volunteers during LREC.

We would like to thank the Japan National Tourism Organization (JNTO) which supported us during LREC’s venue selection process, including site visit to choose the best place for LREC in Japan.

Lastly, we would like to thank Mr. Manmatsu Hayashi for his great effort to make LREC success. The word “impossible” couldn't be found in his dictionary.

We are ready to welcoming you with *omotenashi*, our traditional spirit of hospitality.

Enjoy LREC2018 in Miyazaki!

Table of Contents

Session O1 - Machine Translation & Evaluation

9th May 2018, 11:35

Chair person: **Bente Maegaard** Oral Session

Augmenting Librispeech with French Translations: A Multimodal Corpus for Direct Speech Translation Evaluation

Ali Can Kocabiyikoglu, Laurent Besacier and Olivier Kraif

Recent works in spoken language translation (SLT) have attempted to build end-to-end speech-to-text translation without using source language transcription during learning or decoding. However, while large quantities of parallel texts (such as Europarl, OpenSubtitles) are available for training machine translation systems, there are no large (> 100h) and open source parallel corpora that include speech in a source language aligned to text in a target language. This paper tries to fill this gap by augmenting an existing (monolingual) corpus: LibriSpeech. This corpus, used for automatic speech recognition, is derived from read audiobooks from the LibriVox project, and has been carefully segmented and aligned. After gathering French e-books corresponding to the English audio-books from LibriSpeech, we align speech segments at the sentence level with their respective translations and obtain 236h of usable parallel data. This paper presents the details of the processing as well as a manual evaluation conducted on a small subset of the corpus. This evaluation shows that the automatic alignments scores are reasonably correlated with the human judgments of the bilingual alignment quality. We believe that this corpus (which is made available online) is useful for replicable experiments in direct speech translation or more general spoken language translation experiments.

Evaluating Domain Adaptation for Machine Translation Across Scenarios

Thierry Etchegoyhen, Anna Fernández Torné, Andoni Azpeitia, Eva Martínez Garcia and Anna Matamala

We present an evaluation of the benefits of domain adaptation for machine translation, on three separate domains and language pairs, with varying degrees of domain specificity and amounts of available training data. Domain-adapted statistical and neural machine translation systems are compared to each other and to generic online systems, thus providing an evaluation of the main options in terms of machine translation. Alongside automated translation metrics, we present experimental results involving professional translators, in terms of quality assessment, subjective evaluations of the task and post-editing productivity measurements. The results we present quantify the clear advantages of domain adaptation for machine translation, with marked impacts for domains with higher specificity. Additionally, the results of the experiments show domain-adapted neural machine translation systems to be the optimal choice overall.

Upping the Ante: Towards a Better Benchmark for Chinese-to-English Machine Translation

Christian Hadiwinoto and Hwee Tou Ng

There are many machine translation (MT) papers that propose novel approaches and show improvements over their self-defined baselines. The experimental setting in each paper often differs from one another. As such, it is hard to determine if a proposed approach is really useful and advances the state of the art. Chinese-to-English translation is a common translation direction in MT papers, although there is not one widely accepted experimental setting in Chinese-to-English MT. Our goal in this paper is to propose a benchmark in evaluation setup for Chinese-to-English machine translation, such that the effectiveness of a new proposed MT approach can be directly compared to previous approaches. Towards this end, we also built a highly competitive state-of-the-art MT system trained on a large-scale training set. Our system outperforms reported results on NIST OpenMT test sets in almost all papers published in major conferences and journals in computational linguistics and artificial intelligence in the past 11 years. We argue that a standardized benchmark on data and performance is important for meaningful comparison.

ESCAPE: a Large-scale Synthetic Corpus for Automatic Post-Editing

Matteo Negri, Marco Turchi, Rajen Chatterjee and Nicola Bertoldi

Training models for the automatic correction of machine-translated text usually relies on data consisting of (source,MT, human post-edit) triplets providing, for each source sentence, examples of translation errors with the corresponding corrections made by a human post-editor. Ideally, a large amount of data of this kind should allow the model to learn reliable correction patterns and effectively apply them at test stage on unseen (source, MT) pairs. In practice, however, their limited availability calls for solutions that also integrate in the training process other sources of knowledge. Along this direction, state-of-the-art results have been recently achieved by systems that, in addition to a limited amount of available training data, exploit artificial corpora that approximate elements of the "gold" training instances with automatic translations. Following this idea, we present eSCAPE, the largest freely-available Synthetic Corpus for Automatic Post-Editing released so far. eSCAPE consists of millions of entries in which the MTelement of the training triplets has been obtained by translating the source side of publicly-available parallel

corpora, and using the target side as an artificial human post-edit.Translations are obtained both with phrase-based and neural models. For each MT paradigm, eSCAPE contains 7.2 million triplets forEnglish–German and 3.3 millions for English–Italian, resulting in a total of 14,4 and 6,6 million instances respectively. The usefulness of eSCAPE is proved through experiments in a general-domain scenario, the most challenging one for automatic post-editing. For both language directions, the models trained on our artificial data always improve MT quality with statistically significant gains.

Evaluating Machine Translation Performance on Chinese Idioms with a Blacklist Method

Yutong Shao, Rico Sennrich, Bonnie Webber and Federico Fancellu

Idiom translation is a challenging problem in machine translation because the meaning of idioms is non-compositional, and a literal (word-by-word) translation is likely to be wrong. In this paper, we focus on evaluating the quality of idiom translation of MT systems. We introduce a new evaluation method based on an idiom-specific blacklist of literal translations, based on the insight that the occurrence of any blacklisted words in the translation output indicates a likely translation error. We introduce a dataset, CIBB (Chinese Idioms Blacklists Bank), and perform an evaluation of a state-of-the-art ChineseEnglish neural MT system. Our evaluation confirms that a sizable number of idioms in our test set are mistranslated (46.1%), that literal translation error is a common error type, and that our blacklist method is effective at identifying literal translation errors.

Session O2 - Semantics & Lexicon (1)

9th May 2018, 11:35

Chair person: **Yoshihiko Hayashi** Oral Session

Network Features Based Co-hyponymy Detection

Abhik Jana and Pawan Goyal

Distinguishing lexical relations has been a long term pursuit in natural language processing (NLP) domain. Recently, in order to detect lexical relations like hypernymy, meronymy, co-hyponymy etc., distributional semantic models are being used extensively in some form or the other. Even though a lot of efforts have been made for detecting hypernymy relation, the problem of co-hyponymy detection has been rarely investigated. In this paper, we are proposing a novel supervised model where various network measures have been utilized to identify co-hyponymy relation with high accuracy performing better or at par with the state-of-the-art models.

Cross-Lingual Generation and Evaluation of a Wide-Coverage Lexical Semantic Resource

Attila Novák and Borbála Novák

Neural word embedding models trained on sizable corpora have proved to be a very efficient means of representing meaning. However, the abstract vectors representing words and phrases in these models are not interpretable for humans by themselves. In this paper we present the Thing Recognizer, a method that assigns explicit symbolic semantic features from a finite list of terms to words present in an embedding model, making the model interpretable for humans and covering the semantic space by a controlled vocabulary of semantic features. We do this in a cross-lingual manner, applying semantic tags taken form lexical resources in one language (English) to the embedding space of another (Hungarian)

Advances in Pre-Training Distributed Word Representations

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch and Armand Joulin

Many Natural Language Processing applications nowadays rely on pre-trained word representations estimated from large text corpora such as news collections, Wikipedia and Web Crawl. In this paper, we show how to train high-quality word vector representations by using a combination of known tricks that are however rarely used together. The main result of our work is the new set of publicly available pre-trained models that outperform the current state of the art by a large margin on a number of tasks.

Integrating Generative Lexicon Event Structures into VerbNet

Susan Windisch Brown, James Pustejovsky, Annie Zaenen and Martha Palmer

This paper focuses on specific changes to the semantic representations associated with classes of verbs in the English lexical resource VerbNet. The new form has been restricted to first-order representations to simplify use by and integration with planners. More significantly, the modifications incorporate the Generative Lexicon's event structure, with temporal ordering of subevents associated with explicit predications over the verb's arguments. These changes allow for greater flexibility in representing complex events, for a more consistent treatment of the oppositions inherent in change-of-state classes, and for a more nuanced portrayal of the Agent's role.

FontLex: A Typographical Lexicon based on Affective Associations

Tugba Kulahcioglu and Gerard De Melo

The task of selecting suitable fonts for a given text is non-trivial, as tens of thousands of fonts are available, and the choice of font has been shown to affect the perception of the text as well as of the author or of the brand being advertized. Aiming to support the development of font recommendation tools, we create a typographical lexicon providing associations between words and fonts. We achieve this by means of affective evocations, making use of font–emotion and word–emotion relationships. For this purpose, we first determine font vectors for a set of ten emotion attributes, based on word similarities and antonymy information. We evaluate these associations through a user study via Mechanical Turk, which, for eight of the ten emotions, shows a strong user preference towards the fonts that are found to be congruent by our predicted data. Subsequently, this data is used to calculate font vectors for specific words, by relying on the emotion associations of a given word. This leads to a set of font associations for 6.4K words. We again evaluate the resulting dataset using Mechanical Turk, on 25 randomly sampled words. For the majority of these words, the responses indicate that fonts with strong associations are preferred, and for all except 2 words, fonts with weak associations are dispreferred. Finally, we further extend the dataset using synonyms of font attributes and emotion names. The resulting FontLex resource provides mappings between 6.7K words and 200 fonts.

Session O3 - Corpus Annotation & Tagging

9th May 2018, 11:35

Chair person: **Tomaž Erjavec** Oral Session

Multi-layer Annotation of the Rigveda

Oliver Hellwig, Heinrich Hettrich, Ashutosh Modi and Manfred Pinkal

The paper introduces a multi-level annotation of the Rigveda, a fundamental Sanskrit text composed in the 2. millenium BCE that is important for South-Asian and Indo-European linguistics, as well as Cultural Studies. We describe the individual annotation levels, including phonetics, morphology, lexicon, and syntax, and show how these different levels of annotation are merged to create a novel annotated corpus of Vedic Sanskrit. Vedic Sanskrit is a complex, but computationally under-resourced language. Therefore, creating this resource required considerable domain adaptation of existing computational tools, which is discussed in this paper. Because parts of the annotations are selective, we propose a bi-directional LSTM based sequential model to supplement missing verb-argument links.

The Natural Stories Corpus

Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi and Evelina Fedorenko

It is now a common practice to compare models of human language processing by comparing how well they predict behavioral and neural measures of processing difficulty, such as reading times, on corpora of rich naturalistic linguistic materials. However, many of these corpora, which are based on naturally-occurring text, do not contain many of the low-frequency syntactic constructions that are often required to distinguish between processing theories. Here we describe a new corpus consisting of English texts edited to contain many low-frequency syntactic constructions while still sounding fluent to native speakers. The corpus is annotated with hand-corrected Penn Treebank-style parse trees and includes self-paced reading time data and aligned audio recordings. Here we give an overview of the content of the corpus and release the data.

Semi-automatic Korean FrameNet Annotation over KAIST Treebank

Younggyun Hahm, Jiseong Kim, Sunggoo Kwon and KEY-SUN CHOI

This paper describes a project for constructing FrameNet annotations in Korean over the KAIST treebank corpus to scale up the Korean FrameNet resource. Annotating FrameNet over raw sentences is an expensive and complex task, because of which we have designed this project using a semi-automatic annotation approach. This paper describes the approach and its expected results. As a first step, we built a lexical database of the Korean FrameNet, and used it to learn the model for automatic annotation. Its current scope, status, and limitations are discussed in this paper.

Handling Normalization Issues for Part-of-Speech Tagging of Online Conversational Text

Géraldine Damnati, Jérémy Auguste, Alexis Nasr, Delphine Charlet, Johannes Heinecke and Frédéric Béchet

For the purpose of POS tagging noisy user-generated text, should normalization be handled as a preliminary task or is it possible to handle misspelled words directly in the POS tagging model? We propose in this paper a combined approach where some errors are normalized before tagging, while a Gated Recurrent Unit deep neural network based tagger handles the remaining errors. Word embeddings are trained on a large corpus in order to address both normalization and POS tagging. Experiments are run on Contact Center chat conversations, a particular type of formal Computer Mediated Communication data.

Multi-Dialect Arabic POS Tagging: A CRF Approach

Kareem Darwish, Hamdy Mubarak, Ahmed Abdelali, Mohamed Eldesouki, Younes Samih, Randah Alharbi, Mohammed Attia, Walid Magdy and Laura Kallmeyer

This paper introduces a new dataset of POS-tagged Arabic tweets in four major dialects along with tagging guidelines. The data, which we are releasing publicly, includes tweets in Egyptian, Levantine, Gulf, and Maghrebi, with 350 tweets for each dialect with appropriate train/test/development splits for 5-fold cross validation. We use a Conditional Random Fields (CRF) sequence labeler to train POS taggers for each dialect and examine the effect of cross and joint dialect training, and give benchmark results for the datasets. Using clitic n-grams, clitic metatypes, and stem templates as features, we were able to train a joint model that can correctly tag four different dialects with an average accuracy of 89.3%.

Session O4 - Dialogue

9th May 2018, 11:35

Chair person: **Anna Rumshinsky** Oral Session

A Corpus for Modeling Word Importance in Spoken Dialogue Transcripts

Sushant Kafle and Matt Huenerfauth

Motivated by a project to create a system for people who are deaf or hard-of-hearing that would use automatic speech recognition (ASR) to produce real-time text captions of spoken English during in-person meetings with hearing individuals, we have augmented a transcript of the Switchboard conversational dialogue corpus with an overlay of word-importance annotations, with a numeric score for each word, to indicate its importance to the meaning of each dialogue turn. Further, we demonstrate the utility of this corpus by training an automatic word importance labeling model; our best performing model has an F-score of 0.60 in an ordinal 6-class word-importance classification task with an agreement (concordance correlation coefficient) of 0.839 with the human annotators (agreement score between annotators is 0.89). Finally, we discuss our intended future applications of this resource, particularly for the task of evaluating ASR performance, i.e. creating metrics that predict ASR-output caption text usability for DHH users better thanWord Error Rate (WER).

Dialogue Structure Annotation for Multi-Floor Interaction

David Traum, Cassidy Henry, Stephanie Lukin, Ron Artstein, Felix Gervits, Kimberly Pollard, Claire Bonial, Su Lei, Clare Voss, Matthew Marge, Cory Hayes and Susan Hill

We present an annotation scheme for meso-level dialogue structure, specifically designed for multi-floor dialogue. The scheme includes a transaction unit that clusters utterances from multiple participants and floors into units according to realization of an initiator's intent, and relations between individual utterances within the unit. We apply this scheme to annotate a corpus of multi-floor human-robot interaction dialogues. We examine the patterns of structure observed in these dialogues and present inter-annotator statistics and relative frequencies of types of relations and transaction units. Finally, some example applications of these annotations are introduced.

Effects of Gender Stereotypes on Trust and Likability in Spoken Human-Robot Interaction

Matthias Kraus, Johannes Kraus, Martin Baumann and Wolfgang Minker

As robots enter more and more areas of everyday life, it becomes necessary for them to interact in an understandable and trustworthy way. In many regards this requires a human-like interaction pattern. This research investigates the influence of gender stereotypes on trust and likability of humanoid robots. In this endeavor, explicit (name and voice) and implicit gender (personality) of robots have been manipulated along with the stereotypicality of a task. 40 participants interacted with a NAO robot to gain feedback on a task they were working on and rated the perception of the robot cooperation partner. While no gender stereotypes were found for the explicit gender, implicit gender showed a strong effect on trust and likability in the stereotypical male task. Participants trusted the male robot more and rated it as more reliable and competent than the female personality robot, while the female robot was perceived as more likable. These findings indicate that for gender stereotypes in robot interaction a differentiation between explicit and implicit stereotypical features have to be drawn and that the task context needs consideration. Future research may look into situational variables that drive stereotypification in human-robot interaction.

A Multimodal Corpus for Mutual Gaze and Joint Attention in Multiparty Situated Interaction

Dimosthenis Kontogiorgos, Vanya Avramova, Simon Alexandersson, Patrik Jonell, Catharine Oertel, Jonas Beskow, Gabriel Skantze and Joakim Gustafson

In this paper we present a corpus of multiparty situated interaction where participants collaborated on moving virtual objects on a large touch screen. A moderator facilitated the discussion and directed the interaction. The corpus contains recordings of a variety of multimodal data, in that we captured speech, eye gaze and gesture data using a multisensory setup (wearable eye trackers, motion capture and audio/video). Furthermore, in the description of the multimodal corpus, we investigate four different types of social gaze: referential gaze, joint attention, mutual gaze and gaze aversion by both perspectives of a speaker and a listener. We annotated the groups' object references during object manipulation tasks and analysed the group's proportional referential eye-gaze with regards to the referent object. When investigating the distributions of gaze during and before referring expressions we could corroborate the differences in time between speakers' and listeners' eye gaze found in earlier studies. This corpus is of particular interest to researchers who are interested in social eye-gaze patterns in turn-taking and referring language in situated multi-party interaction.

Improving Dialogue Act Classification for Spontaneous Arabic Speech and Instant Messages at Utterance Level

AbdelRahim Elmadany, Sherif Abdou and Mervat Gheith

The ability to model and automatically detect dialogue act is an important step toward understanding spontaneous speech and Instant Messages. However, it has been difficult to infer a dialogue act from a surface utterance because it highly depends on the context of the utterance and speaker linguistic knowledge; especially in Arabic dialects. This paper proposes a statistical dialogue analysis model to recognize utterance's dialogue acts using a multi-classes hierarchical structure. The model can automatically acquire probabilistic discourse knowledge from a dialogue corpus were collected and annotated manually from multi-genre Egyptian call-centers. Extensive experiments were conducted using Support Vector Machines classifier to evaluate the system performance. The results attained in the term of average F-measure scores of 0.912; showed that the proposed approach has moderately improved F-measure by approximately 20%.

Session P1 - Anaphora, Coreference

9th May 2018, 11:35

Chair person: **Scott Piao** Poster Session

Coreference Resolution in FreeLing 4.0

Montserrat Marimon, Lluís Padró and Jordi Turmo

This paper presents the integration of RelaxCor into FreeLing. RelaxCor is a coreference resolution system based on constraint satisfaction that ranked second in CoNLL-2011 shared task. FreeLing is an open-source library for NLP with more than fifteen years of existence and a widespread user community. We present the difficulties found in porting RelaxCor from a shared task scenario to a production enviroment, as well as the solutions devised. We present two strategies for this integration and a rough evaluation of the obtained results.

BASHI: A Corpus of Wall Street Journal Articles Annotated with Bridging Links

Ina Roesiger

This paper presents a corpus resource for the anaphoric phenomenon of bridging, named BASHI. The corpus consisting of 50 Wall Street Journal (WSJ) articles adds bridging anaphors and their antecedents to the other gold annotations that have been created as part of the OntoNotes project (Weischedel et al. 2011). Bridging anaphors are context-dependent expressions that do not refer to the same entity as their antecedent, but to a related entity. Bridging resolution is an under-researched area of NLP, where the lack of annotated training data makes the application of statistical models difficult. Thus, we believe that the corpus is a valuable resource for researchers interested in anaphoric phenomena going beyond coreference, as it can be combined with other corpora to create a larger corpus resource. The corpus contains 57,709 tokens and 459 bridging pairs and is available for download in an offset-based format and a CoNLL-12 style bridging column that can be merged with the other annotation layers in OntoNotes. The paper also reviews previous annotation efforts and different definitions of bridging and reports challenges with respect to the bridging annotation.

SACR: A Drag-and-Drop Based Tool for Coreference Annotation

Bruno Oberle

This paper introduces SACR, an easy-to-use coreference chain annotation tool, which is used to annotate large corpora for Natural Language Processing applications. Coreference annotation is usually considered as costly both in terms of time and human resources. So, in order to find the easiest annotation

strategy, we will first of all compare several annotation schemes implemented in existing tools. Since interface ergonomics is also an important part of our research, we then focus on identifying the most helpful features to reduce the strain for annotators. In the next section of the paper, we present SACR in details. This tool has been developped specifically for coreference annotation, and its intuitive user interface has been designed to facilitate and speed up the annotation process, making SACR equally suited for students, occasional and non-technical users. In order to create coreference chains, elements are selected by clicking on the corresponding tokens. Coreference relations are then created by drag-and-dropping expressions one over the other. Finally, color frames around marked expressions help the user to visualize both marked expressions and their relations. SACR is open source, distributed under the terms of the Mozilla Public License, version 2.0, and freely available online.

Deep Neural Networks for Coreference Resolution for Polish

Bartłomiej Nitoń, Paweł Morawiecki and Maciej Ogrodniczuk

The paper presents several configurations of deep neural networks aimed at the task of coreference resolution for Polish. Starting with the basic feature set and standard word embedding vector size we examine the setting with larger vectors, more extensive sets of mention features, increased number of negative examples, Siamese network architecture and a global mention connection algorithm. The highest results are achieved by the system combining our best deep neural architecture with the sieve-based approach – the cascade of rule-based coreference resolvers ordered from most to least precise. All systems are evaluated on the data of the Polish Coreference Corpus featuring 540K tokens and 180K mentions. The best variant improves the state of the art for Polish by 0.53 F1 points, reaching 81.23 points of the CoNLL metric.

SzegedKoref: A Hungarian Coreference Corpus

Veronika Vincze, Klára Hegedűs, Alex Sliz-Nagy and Richárd Farkas

In this paper we introduce SzegedKoref, a Hungarian corpus in which coreference relations are manually annotated. For annotation, we selected some texts of Szeged Treebank, the biggest treebank of Hungarian with manual annotation at several linguistic layers. The corpus contains approximately 55,000 tokens and 4000 sentences. Due to its size, the corpus can be exploited in training and testing machine learning based coreference resolution systems, which we would like to implement in the near future. We present the annotated texts, we describe the annotated categories of anaphoric relations, we report on the annotation process and we offer several examples of each annotated category. Two linguistic phenomena – phonologically empty pronouns and pronouns referring to subordinate clauses – are important characteristics of Hungarian coreference relations. In our paper, we also discuss both of them.

A Corpus to Learn Refer-to-as Relations for Nominals

Wasi Ahmad and Kai-Wei Chang

Continuous representations for words or phrases, trained on large unlabeled corpora are proved very useful for many natural language processing tasks. While these vector representations capture many fine-grained syntactic and semantic regularities among words or phrases, it often lacks coreferential information which is useful for many downstream tasks like information extraction, text summarization etc. In this paper, we argue that good word and phrase embeddings should contain information for identifying refer-to-as relationship and construct a corpus from Wikipedia to generate coreferential neural embeddings for nominals. The term nominal refers to a word or a group of words that functions like a noun phrase. In addition, we use coreference resolution as a proxy to evaluate the learned neural embeddings for noun phrases. To simplify the evaluation procedure, we design a coreferential phrase prediction task where the learned nominal embeddings are used to predict which candidate nominals can be referred to a target nominal. We further describe how to construct an evaluation dataset for such task from well known OntoNotes corpus and demonstrate encouraging baseline results.

Sanaphor++: Combining Deep Neural Networks with Semantics for Coreference Resolution

Julien Plu, Roman Prokofyev, Alberto Tonon, Philippe Cudré-Mauroux, Djellel Eddine Difallah, Raphael Troncy and Giuseppe Rizzo

Coreference resolution has always been a challenging task in Natural Language Processing. Machine learning and semantic techniques have improved the state of the art over the time, though since a few years, the biggest step forward has been made using deep neural networks. In this paper, we describe Sanaphor++, which is an improvement of a top-level deep neural network system for coreference resolution—namely Stanford deep-coref—through the addition of semantic features. The goal of Sanaphor++ is to improve the clustering part of the coreference resolution in order to know if two clusters have to be merged or not once the pairs of mentions have been identified. We evaluate our model over the CoNLL 2012 Shared Task dataset and compare it with the state-of-the-art system (Stanford deep-coref) where we demonstrated an average gain of 1.13\% of the average F1 score.

ANCOR-AS: Enriching the ANCOR Corpus with Syntactic Annotations

Loïc Grobol, Isabelle Tellier, Eric De La Clergerie, Marco Dinarelli and Frédéric Landragin

This paper presents ANCOR-AS, an enriched version of the ANCOR corpus. This version adds syntactic annotations in addition to the existing coreference and speech transcription ones. This corpus is also released in a new TEI-compliant XML format.

ParCorFull: a Parallel Corpus Annotated with Full Coreference

Ekaterina Lapshinova-Koltunski, Christian Hardmeier and Pauline Krielke

In this paper, we describe a parallel corpus annotated with full coreference chains that has been created to address an important problem that machine translation and other multilingual natural language processing (NLP) technologies face – translation of coreference across languages. Recent research in multilingual coreference and automatic pronoun translation has led to important insights into the problem and some promising results. However, its scope has been restricted to pronouns, whereas the phenomenon is not limited to anaphoric pronouns. Our corpus contains parallel texts for the language pair English-German, two major European languages. Despite being typologically very close, these languages still have systemic differences in the realisation of coreference, and thus pose problems for multilingual coreference resolution and machine translation. Our parallel corpus with full annotation of coreference will be a valuable resource with a variety of uses not only for NLP applications, but also for contrastive linguists and researchers in translation studies. This resource supports research on the mechanisms involved in coreference translation in order to develop a better understanding of the phenomenon. The corpus is available from the LINDAT repository at http://hdl.handle.net/11372/LRT-2614.

Session: Session P2 - Collaborative Resource Construction & Crowdsourcing

9th May 2018, 11:35

Chair person: **Asad Sayeed** Poster Session

An Application for Building a Polish Telephone Speech Corpus

Bartosz Ziółko, Piotr Żelasko, Ireneusz Gawlik, Tomasz Pędzimąż and Tomasz Jadczyk

The paper presents our approach towards building a tool for speech corpus collection of a specific domain content. We describe our iterative approach to the development of this tool, with focus on the most problematic issues at each working stage. Our latest version synchronizes VoIP call management and recording with a web application providing content. The tool was already used and applied for Polish to gather 63 hours of automatically annotated recordings across several domains. Amongst them, we obtained a continuous speech corpus designed with an emphasis on optimal phonetic diversification in relation to the phonetically balanced National Corpus of Polish. We evaluate the usefulness of this data against the GlobalPhone corpus in the task of training an acoustic model for a telephone speech ASR system and show that the model trained on our balanced corpus achieves significantly lower WER in two grammar-based speech recognition tasks - street names and public transport routes numbers.

CPJD Corpus: Crowdsourced Parallel Speech Corpus of Japanese Dialects

Shinnosuke Takamichi and Hiroshi Saruwatari

Public parallel corpora of dialects can accelerate related studies such as spoken language processing. Various corpora have been collected using a well-equipped recording environment, such as voice recording in an anechoic room. However, due to geographical and expense issues, it is impossible to use such a perfect recording environment for collecting all existing dialects. To address this problem, we used web-based recording and crowdsourcing platforms to construct a crowdsourced parallel speech corpus of Japanese dialects (CPJD corpus) including parallel text and speech data of 21 Japanese dialects. We recruited native dialect speakers on the crowdsourcing platform, and the hired speakers recorded their dialect speech using their personal computer or smartphone in their homes. This paper shows the results of the data collection and analyzes the audio data in terms of the signal-to-noise ratio and mispronunciations.

Korean L2 Vocabulary Prediction: Can a Large Annotated Corpus be Used to Train Better Models for Predicting Unknown Words?

Kevin Yancey and Yves Lepage

Vocabulary knowledge prediction is an important task in lexical text simplification for foreign language learners (L2 learners). However, previously studied methods that use hand-crafted rules based on one or two word features have had limited success. A recent study hypothesized that a supervised learning classifier trained on a large annotated corpus of words unknown by L2 learners may yield better results. Our study crowdsourced the production of such a corpus for Korean, now consisting of 2,385 annotated passages contributed by 357 distinct L2 learners. Our preliminary evaluation of models trained on this corpus show favorable results, thus confirming the hypothesis. In this paper, we describe our methodology for building this resource in detail

and analyze its results so that it can be duplicated for other languages. We also present our preliminary evaluation of models trained on this annotated corpus, the best of which recalls 80% of unknown words with 71% precision. We make our annotation data available.

Crowdsourcing-based Annotation of the Accounting Registers of the Italian Comedy

Adeline Granet, Benjamin Hervy, Geoffrey Roman-Jimenez, Marouane Hachicha, Emmanuel Morin, Harold Mouchère, Solen Quiniou, Guillaume Raschia, Françoise Rubellin and Christian Viard-Gaudin

In this paper, we present a double annotation system for new handwritten historical documents. We have 25,250 pages of registers of the Italian Comedy of the 18th century containing a great variety and amount of information. A crowdsourcing platform has been set up in order to perform labeling and transcription of the documents. The main purpose is to grasp budget data from the all 18th century and to create a dedicated database for the domain's experts. In order to improve, help and accelerate the process, a parallel system has been designed to automatically process information. We focus on the titles field, segmenting them into lines and checking candidate transcripts. We have collected a base of 971 title lines.

FEIDEGGER: A Multi-modal Corpus of Fashion Images and Descriptions in German

Leonidas Lefakis, Alan Akbik and Roland Vollgraf

The availability of multi-modal datasets that pair images and textual descriptions of their content has been a crucial driver in progress of various text-image tasks such as automatic captioning and text-to-image retrieval. In this paper, we present FEIDEGGER, a new multi-modal corpus that focuses specifically on the domain of fashion items and their visual descriptions in German. We argue that such narrow-domain multi-modality presents a unique set of challenges such as fine-grained image distinctions and domain-specific language, and release this dataset to the research community to enable study of these challenges. This paper illustrates our crowdsourcing strategy to acquire the textual descriptions, gives an overview over the \dataset~dataset, and discusses possible use cases.

Toward a Lightweight Solution for Less-resourced Languages: Creating a POS Tagger for Alsatian Using Voluntary Crowdsourcing

Alice Millour and Karën Fort

We present here the results of an experiment aiming at crowdsourcing part-of-speech annotations for a less-resourced French regional language, Alsatian. We used for this purpose a specifically-developed slightly gamified platform, Bisame. It allowed us to gather annotations on a variety of corpora covering some of the language dialectal variations. The quality of the annotations, which reach an averaged F-measure of 93%, enabled us to train a first tagger for Alsatian that is nearly 84% accurate. The platform as well as the produced annotations and tagger are freely available. The platform can easily be adapted to other languages, thus providing a solution to (some of) the less-resourced languages issue.

Crowdsourced Corpus of Sentence Simplification with Core Vocabulary

Akihiro Katsuta and Kazuhide Yamamoto

We present a new Japanese crowdsourced data set of simplified sentences created from more complex ones. Our simplicity standard involves all rewritable words in the simplified sentences being drawn from a core vocabulary of 2,000 words. Our simplified corpus is a collection of complex sentences from Japanese textbooks and reference books together with simplified sentences generated by humans, paired with data on how the complex sentences were paraphrased. The corpus contains a total of 15,000 sentences, in both complex and simple versions. In addition, we investigate the differences in the simplification operations used by each annotator. The aim is to understand whether a crowdsourced complex-simple parallel corpus is an appropriate data source for automated simplification by machine learning. The results, that there was a high level of agreement between the annotators building the data set. So, we believe that this corpus is a good quality data set for machine learning for simplification. We therefore plan to expand the scale of the simplified corpus in the future.

A Multilingual Wikified Data Set of Educational Material

Iris Hendrickx, Eirini Takoulidou, Thanasis Naskos, Katia Lida Kermanidis, Vilelmini Sosoni, Hugo De Vos, Maria Stasimioti, Menno Van Zaanen, Panayota Georgakopoulou, Valia Kordoni, Maja Popovic, Markus Egg and Antal Van den Bosch

We present a parallel wikified data set of parallel texts in eleven language pairs from the educational domain. English sentences are lined up to sentences in eleven other languages (BG, CS, DE, EL, HR, IT, NL, PL, PT, RU, ZH) where names and noun phrases (entities) are manually annotated and linked to their respective Wikipedia pages. For every linked entity in English, the corresponding term or phrase in the target language is also marked and linked to its Wikipedia page in that language. The annotation process was performed via crowdsourcing. In this paper we present the task, annotation process, the encountered

difficulties with crowdsourcing for complex annotation, and the data set in more detail. We demonstrate the usage of the data set for Wikification evaluation. This data set is valuable as it constitutes a rich resource consisting of annotated data of English text linked to translations in eleven languages including several languages such as Bulgarian and Greek for which not many LT resources are available.

Using Crowd Agreement for Wordnet Localization

Amarsanaa Ganbold, Altangerel Chagnaa and Gábor Bella

Building a wordnet from scratch is a huge task, especially for languages less equipped with pre-existing lexical resources such as thesauri or bilingual dictionaries. We address the issue of costliness of human supervision through crowdsourcing that offers a good trade-off between quality of output and speed of progress. In this paper, we demonstrate a two-phase crowdsourcing workflow that consists of a synset localization step followed by a validation step. Validation is performed using the inter-rater agreement metrics Fleiss' kappa and Krippendorf's alpha, which allow us to estimate the precision of the result, as well as to set a balance between precision and recall. In our experiment, 947 synsets were localized from English to Mongolian and evaluated through crowdsourcing with the precision of 0.74.

Translation Crowdsourcing: Creating a Multilingual Corpus of Online Educational Content

Vilelmini Sosoni, Katia Lida Kermanidis, Maria Stasimioti, Thanasis Naskos, Eirini Takoulidou, Menno Van Zaanen, Sheila Castilho, Panayota Georgakopoulou, Valia Kordoni and Markus Egg

The present work describes a multilingual corpus of online content in the educational domain, i.e. Massive Open Online Course material, ranging from course forum text to subtitles of online video lectures, that has been developed via large-scale crowdsourcing. The English source text is manually translated into 11 European and BRIC languages using the CrowdFlower platform. During the process several challenges arose which mainly involved the in-domain text genre, the large text volume, the idiosyncrasies of each target language, the limitations of the crowdsourcing platform, as well as the quality assurance and workflow issues of the crowdsourcing process. The corpus constitutes a product of the EU-funded TraMOOC project and is utilised in the project in order to train, tune and test machine translation engines.

Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing

Yo Ehara

We introduce a freely available dataset for analyzing the English vocabulary of English-as-a-second language (ESL) learners. While ESL vocabulary tests have been extensively studied, few of the results have been made public. This is probably because 1) most of the tests are used to grade test takers, i.e., placement tests; thus, they are treated as private information that should not be leaked, and 2) the primary focus of most language-educators is how to measure their students' ESL vocabulary, rather than the test results of the other test takers. However, to build and evaluate systems to support language learners, we need a dataset that records the learners' vocabulary. Our dataset meets this need. It contains the results of the vocabulary size test, a well-studied English vocabulary test, by one hundred test takers hired via crowdsourcing. Unlike high-stakes testing, the test takers of our dataset were not motivated to cheat on the tests to obtain high scores. This setting is similar to that of typical language-learning support systems. Brief test-theory analysis on the dataset showed an excellent test reliability of 0.91 (Chronbach's alpha). Analysis using item response theory also indicates that the test is reliable and successfully measures the vocabulary ability of language learners. We also measured how well the responses from the learners can be predicted with high accuracy using machine-learning methods.

Session P3 - Information Extraction, Information Retrieval, Text Analytics (1)

9th May 2018, 11:35

Chair person: **Hikaru Yokono** Poster Session

Chinese Relation Classification using Long Short Term Memory Networks

Linrui Zhang and Dan Moldovan

Relation classification is the task to predict semantic relations between pairs of entities in a given text. In this paper, a novel Long Short Term Memory Network (LSTM)-based approach is proposed to extract relations between entities in Chinese text. The shortest dependency path (SDP) between two entities, together with the various selected features in the path, are rst extracted, and then used as input of an LSTM model to predict the relation between them. The performance of the system was evaluated on the ACE 2005 Multilingual Training Corpus, and achieved a state-of-the-art F-measure of 87.87% on six general type relations and 83.40% on eighteen subtype relations in this corpus.

The UIR Uncertainty Corpus for Chinese: Annotating Chinese Microblog Corpus for Uncertainty Identification from Social Media

Binyang Li, Jun Xiang, Le Chen, Xu Han, Xiaoyan Yu, Ruifeng Xu, Tengjiao Wang and Kam-Fai Wong

Uncertainty identification is an important semantic processing task, which is critical to the quality of information in terms of factuality in many NLP techniques and applications, such as question answering, information extraction, and so on. Especially in social media, the factuality becomes a primary concern, because the social media texts are usually written wildly. The lack of open uncertainty corpus for Chinese social media contexts bring limitations for many social media oriented applications. In this work, we present the first open uncertainty corpus of microblogs in Chinese, namely, the UIR Uncertainty Corpus (UUC). At current stage, we annontated 40,168 Chinese microblogs from Sina Microblog. The schema of CoNLL 2010 have been adapted, where the corpus contains annotations at each microblog level for uncertainty and 6 sub-classes with 11,071 microblogs under uncertainty. To adapt to the characeristics of social media, we identify the uncertainty based on the contextual uncertan semantics rather than the traditional cue-phrases, and the sub-class could provide more information for research on handing uncertainty in social media texts. The Kappa value indicated that our annotation results were substantially reliable.

EventWiki: A Knowledge Base of Major Events

Tao Ge, Lei Cui, Baobao Chang, Zhifang Sui, Furu Wei and Ming Zhou

This paper introduces a new resource called EventWiki which is, to the best of our knowledge, the first knowledge base resource of major events. In contrast to most existing knowledge bases that focus on static entities such as people, locations and organizations, our EventWiki concentrate on major events, in which all entries in EventWiki are important events in mankind history. We demonstrate that EventWiki is a very useful resource for information extraction regarding events in Natural Language Processing (NLP), knowledge inference and automatic knowledge base construction.

Annotating Spin in Biomedical Scientific Publications : the case of Random Controlled Trials (RCTs)

Anna Koroleva and Patrick Paroubek

In this paper we report on the collection in the context of the MIROR project of a corpus of biomedical articles for the task of automatic detection of inadequate claims (spin), which to our knowledge has never been addressed before. We present the manual annotation model and its annotation guidelines and describe the planned machine learning experiments and evaluations.

Visualization of the occurrence trend of infectious diseases using Twitter

Ryusei Matsumoto, Minoru Yoshida, Kazuyuki Matsumoto, Hironobu Matsuda and Kenji Kita

We propose a system for visualizing the epidemics of infectious diseases. We apply factuality analysis both to disease detection and location estimation for accurate visualization. We tested our methods for several infectious diseases, and show that our method performs well on various diseases.

Reusable workflows for gender prediction

Matej Martinc and Senja Pollak

This paper presents a system for author profiling (AP) modeling that reduces the complexity and time of building a sophisticated model for a number of different AP tasks. The system is implemented in a cloud-based visual programming platform ClowdFlows and is publicly available to a wider audience. In the platform, we also implemented our already existing state of the art gender prediction model and tested it on a number of cross-genre tasks. The results show that the implemented model, which was trained on tweets, achieves results comparable to state of the art models for cross-genre gender prediction. There is however a noticeable decrease in accuracy when the genre of a test set is different from the genre of the train set.

Knowing the Author by the Company His Words Keep

Armin Hoenen and Niko Schenk

In this paper, we analyze relationships between word pairs and evaluate their idiosyncratic properties in the applied context of authorship attribution. Specifically, on three literary corpora we optimize word pair features for information gain which reflect word similarity as measured by word embeddings. We analyze the quality of the most informative features in terms of word type relation (a comparison of different constellations of function and content words), similarity, and relatedness. Results point to the extraordinary role of function words within the authorship attribution task being extended to their pairwise relational patterns. Similarity of content words is likewise among the most informative features. From a cognitive perspective, we conclude that both relationship types reflect short distance connections in the human brain, which is highly indicative of an individual writing style.

Towards a Gold Standard Corpus for Variable Detection and Linking in Social Science Publications

Andrea Zielinski and Peter Mutschke

this paper, we describe our effort to create a new corpus for the evaluation of detecting and linking so-called survey variables in social science publications (e.g. "Do you believe in Heaven?"). The task is to recognize survey variable mentions in a given text, disambiguate them, and link them to the corresponding variable within a knowledge base. Since there are generally hundreds of candidates to link to and due to the wide variety of forms they can take, this is a challenging task within NLP. The contribution of our work is the first gold standard corpus for the variable detection task. We describe the annotation guidelines and the annotation process. The produced corpus is multilingual - German and English - and includes manually curated word and phrase alignments. Moreover, it includes text samples that could not be assigned to any variables, denoted as negative examples. Based on the new dataset, we conduct an evaluation of several state-of-the-art text classification and textual similarity methods. The annotated corpus is made available along with an open-source baseline system for variable mention identification and linking.

KRAUTS: A German Temporally Annotated News Corpus

Jannik Strötgen, Anne-Lyse Minard, Lukas Lange, Manuela Speranza and Bernardo Magnini

In recent years, temporal tagging, i.e., the extraction and normalization of temporal expressions, has become a vivid research area. Several tools have been made available, and new strategies have been developed. Due to domain-specific challenges, evaluations of new methods should be performed on diverse text types. Despite significant efforts towards multilinguality in the context of temporal tagging, for all languages except English, annotated corpora exist only for a single domain. In the case of German, for example, only a narrative-style corpus has been manually annotated so far, thus no evaluations of German temporal tagging performance on news articles can be made. In this paper, we present KRAUTS, a new German temporally annotated corpus containing two subsets of news documents: articles from the daily newspaper Dolomiten and from the weekly newspaper Die Zeit. Overall, the corpus contains 192 documents with 1,140 annotated temporal expressions, and has been made publicly available to further boost research in temporal tagging.

Session P4 - Infrastructural Issues/Large Projects (1)

9th May 2018, 11:35

Chair person: **Denise Di Persio** Poster Session

CogCompNLP: Your Swiss Army Knife for NLP

Daniel Khashabi, Mark Sammons, Ben Zhou, Tom Redman, Christos Christodoulopoulos, Vivek Srikumar, Nickolas Rizzolo, Lev Ratinov, Guanheng Luo, Quang Do, Chen-Tse Tsai, Subhro Roy, Stephen Mayhew, Zhili Feng, John Wieting, Xiaodong Yu, Yangqiu Song, Shashank Gupta, Shyam Upadhyay, Naveen Arivazhagan, Qiang Ning, Shaoshi Ling and Dan Roth

Implementing a Natural Language Processing (NLP) system requires considerable engineering effort: creating data-structures to represent language constructs; reading corpora annotations into these data-structures; applying off-the-shelf NLP tools to augment the text representation; extracting features and training machine learning components; conducting experiments and computing performance statistics; and creating the end-user application that integrates the implemented components. While there are several widely used NLP libraries, each provides only partial coverage of these various tasks. We present our library COGCOMPNLP which simplifies the process of design and development of NLP applications by providing modules to address different challenges: we provide a corpus-reader module that supports popular corpora in the NLP community, a module for various low-level data-structures and operations (such as search over text), a module for feature extraction, and an extensive suite of annotation modules for a wide range of semantic and syntactic tasks. These annotation modules are all integrated in a single system, PIPELINE, which allows users to easily use the annotators with simple direct calls using any JVM-based language, or over a network. The sister project COGCOMPNLPYenables users to access the annotators with a Python interface. We give a detailed account of our system's structure and usage, and where possible, compare it with other established NLP frameworks. We report on the performance, including time and memory statistics, of each component on a selection of well-established datasets. Our system is publicly available for research use and external contributions, at:http://github.com/CogComp/cogcomp-nlp

A Framework for the Needs of Different Types of Users in Multilingual Semantic Enrichment

Jan Nehring and Felix Sasaki

The FREME framework bridges Language Technologies (LT) and Linked Data (LD). It establishes workflows between LT

and LD in a well defined, coherent way. FREME addresses common challenges that both researchers and industry face when integrating LT and LD: interoperability, "silo" solutions and the lack of adequate tooling. Usability, reusability and interoperability are often attributes of frameworks and toolkits for LT and LD. In this paper, we take a novel approach: We define user types and user levels and describe how they influence design decisions in a LT and LD processing framework. In this way, we combine research outcomes from various communities: language technology, linked data and software interface engineering. This paper explains the different user types and how FREME addresses the specific needs of each user type. Core attributes of FREME are usability, reusability and interoperability

The LREC Workshops Map

Roberto Bartolini, Sara Goggi, Monica Monachini and Gabriella Pardelli

The aim of this work is to present an overview of the research presented at the LREC workshops over the years 1998-2016 with the aim to shed light on the community represented by workshop participants in terms of country of origin, type of affiliation, gender. There has been also an effort towards the identification of the major topics dealt with as well as of the terminological variations noticed in this time span. Data has been retrieved from the portal of the European Language Resources Association (ELRA) which organizes the conference and the resulting corpus made up of workshops titles and of the related presentations has then been processed using a term extraction tool developed at ILC-CNR.

Preserving Workflow Reproducibility: The RePlay-DH Client as a Tool for Process Documentation

Markus Gärtner, Uli Hahn and Sibylle Hermann

In this paper we present a software tool for elicitation and management of process metadata. It follows our previously published design idea of an assistant for researchers that aims at minimizing the additional effort required for producing a sustainable workflow documentation. With the ever-growing number of linguistic resources available, it also becomes increasingly important to provide proper documentation to make them comparable and to allow meaningful evaluations for specific use cases. The often prevailing practice of post hoc documentation of resource generation or research processes bears the risk of information loss. Not only does detailed documentation of a process aid in achieving reproducibility, it also increases usefulness of the documented work for others as a cornerstone of good scientific practice. Time pressure together with the lack of simple documentation methods leads to workflow documentation in practice being an arduous and often neglected task. Our tool ensures a clean documentation for common workflows in natural language processing and digital humanities. Additionally, it can easily be integrated into existing institutional infrastructures.

The ACoLi CoNLL Libraries: Beyond Tab-Separated Values

Christian Chiarcos and Niko Schenk

We introduce the ACoLi CoNLL libraries, a set of Java archives to facilitate advanced manipulations of corpora annotated in TSV formats, including all members of the CoNLL format family. In particular, we provide means for (i) rule-based re-write operations, (ii) visualization and manual annotation, (iii) merging CoNLL files, and (iv) data base support. The ACoLi CoNLL libraries provide command-line interface to these functionalities. The following aspects are technologically innovative and exceed beyond the state of the art: We support \emph{every} OWPL (one word per line) corpus format with tab-separated columns, whereas most existing tools are specific to one particular CoNLL dialect. We employ established W3C standards for rule-based graph rewriting operations on CoNLL sentences. We provide means for the heuristic, but fully automated merging of CoNLL annotations of the same textual content, in particular for resolving conflicting tokenizations. We demonstrate the usefulness and practicability of our proposed CoNLL libraries on well-established data sets of the Universal Dependency corpus and the Penn Treebank.

What's Wrong, Python? – A Visual Differ and Graph Library for NLP in Python

Balázs Indig, András Simonyi and Noémi Ligeti-Nagy

The correct analysis of the output of a program based on supervised learning is inevitable in order to be able to identify the errors it produced and characterise its error types. This task is fairly difficult without a proper tool, especially if one works with complex data structures such as parse trees or sentence alignments. In this paper, we present a library that allows the user to interactively visualise and compare the output of any program that yields a well-known data format. Our goal is to create a tool granting the total control of the visualisation to the user, including extensions, but also have the common primitives and data-formats implemented for typical cases. We describe the common features of the common NLP tasks from the viewpoint of visualisation in order to specify the essential primitive functions. We enumerate many popular off-the-shelf NLP visualisation programs to compare with our implementation, which unifies all of the profitable features of the existing programs adding extendibility as a crucial feature to them.

Session P5 - Knowledge Discovery/Representation

9th May 2018, 11:35

Chair person: **Dan Tufiș** Poster Session

ScholarGraph:a Chinese Knowledge Graph of Chinese Scholars

Shuo Wang, Zehui Hao, Xiaofeng Meng and Qiuyue Wang

Scholars and their academic information are widely distributed on the Web. Integrating these information and making association between them can play a catalytic role in academic evaluation and research. Since 2008, Web and Mobile Data Management Laboratory (WAMDM) in Renmin University of China began to collect Chinese literatures in more than 20 academic domains, and build a data integration system called ScholarSpace to automatically integrate the relevant chinese academic information from the chinese scholars and science. Focusing on the chinese scholars, ScholarSpace can give you an academic portrait about a chinese scholar with the form of knowledge graph. So the ScholarSpace can be tranformed into a knowledge graph called ScholarGraph. It includes the scholar information such as the affiliation, publications, teacher-student relationship, etc. ScholarGraph is a subset of the whole knowledge graph generated from the ScholarSpace and is published on the web page of WAMDM. ScholarGraph consists of more than 10,000,000 triples, including more than 9,000,000 entities and 6 relations. It can support the search and query about portrait of Chinese scholars and other relevant applications.

Enriching Frame Representations with Distributionally Induced Senses

Stefano Faralli, Alexander Panchenko, Chris Biemann and Simone Paolo Ponzetto

We introduce a new lexical resource that enriches the Framester knowledge graph, which links Framnet, WordNet, VerbNet and other resources, with semantic features from text corpora. These features are extracted from distributionally induced sense inventories and subsequently linked to the manually-constructed frame representations to boost the performance of frame disambiguation in context. Since Framester is a frame-based knowledge graph, which enables full-fledged OWL querying and reasoning, our resource paves the way for the development of novel, deeper semantic-aware applications that could benefit from the combination of knowledge from text and complex symbolic representations of events and participants. Together with the resource we also provide the software we developed for the evaluation in the task of Word Frame Disambiguation (WFD).

An Integrated Formal Representation for Terminological and Lexical Data included in Classification Schemes

Thierry Declerck, Kseniya Egorova and Eileen Schnur

This paper presents our work dealing with a potential application in e-lexicography: the automatized creation of specialized multilingual dictionaries from structured data, which are available in the form of comparable multilingual classification schemes or taxonomies. As starting examples, we use comparable industry classification schemes, which frequently occur in the context of stock exchanges and business reports. Initially, we planned to follow an approach based on cross-taxonomies and cross-languages string mapping to automatically detect candidate multilingual dictionary entries for this specific domain. However, the need to first transform the comparable classification schemes into a shared formal representation language in order to be able to properly align their components before implementing the algorithms for the multilingual lexicon extraction soon became apparent. We opted for the SKOS-XL vocabulary for modelling the multilingual terminological part of the comparable taxonomies and for OntoLex-Lemon for modelling the multilingual lexical entries which can be extracted from the original data. In this paper, we present the suggested modelling architecture, which demonstrates how terminological elements and lexical items can be formally integrated and explicitly cross-linked in the context of the Linguistic Linked Open Data (LLOD).

One event, many representations. Mapping action concepts through visual features.

Alessandro Panunzi, Lorenzo Gregori and Andrea Amelio Ravelli

This paper faces the problem of unifying the representation of actions and events in different semantic resources. The proposed solution exploits the IMAGACT visual component (video scenes that represent physical actions) as the linkage point among resources. By using visual objects, we connected resources responding to different scopes and theoretical frameworks, in which a concept-to-concept mapping appeared difficult to obtain. We provide a brief description of two experiments that exploit IMAGACT videos as a linkage point: an automatic linking with BabelNet, a multilingual semantic network, and a manual linking with Praxicon, a conceptual knowledge base of action. The aim of this work is to integrate data from resources with different level of granularity in order to describe the action semantics from a linguistic, visual and motor point of view.

Tel(s)-Telle(s)-Signs: Highly Accurate Automatic Crosslingual Hypernym Discovery

Ada Wan

We report a highly accurate hypernym discovery heuristic that works on unrestricted texts. This approach leverages morphological cues in French, but given any parallel data and word alignment tool, this proves to be a technique that can work reliably in other languages as well. We tested this method using two French-English corpora of different genres (medical and news) and attained near-perfect accuracy. The key idea is to exploit morphological information in the French trigger phrase 'tel(s)/telle(s)- que' (meaning "such as" in English) to uniquely identify the correct hypernym. This shows to be an inexpensive and effective heuristic also when there are multiple noun phrases preceding the trigger phrase, as in the case of prepositional phrase attachment causing ambiguity in interpretation and hypernym acquisition, indicating that this pattern in French is more informative than its English counterpart.

Session P6 - Opinion Mining / Sentiment Analysis (1)

9th May 2018, 11:35

Chair person: **Cristina Bosco** Poster Session

Disambiguation of Verbal Shifters

Michael Wiegand, Sylvette Loda and Josef Ruppenhofer

Negation is an important contextual phenomenon that needs to be addressed in sentiment analysis. Next to common negation function words, such as "not" or "none", there is also a considerably large class of negation content words, also referred to as shifters, such as the verbs "diminish", "reduce" or "reverse". However, many of these shifters are ambiguous. For instance, "spoil" as in "spoil your chance" reverses the polarity of the positive polar expression "chance" while in "spoil your loved ones", no negation takes place. We present a supervised learning approach to disambiguating verbal shifters. Our approach takes into consideration various features, particularly generalization features.

Bootstrapping Polar-Opposite Emotion Dimensions from Online Reviews

Luwen Huangfu and Mihai Surdeanu

We propose a novel bootstrapping approach for the acquisition of lexicons from unannotated, informal online texts (in our case, Yelp reviews) for polar-opposite emotion dimension values from the Ortony/Clore/Collins model of emotions (e.g., desirable/undesirable). Our approach mitigates the intrinsic problem of limited supervision in bootstrapping with an effective strategy that softly labels unlabeled terms, which are then used to better estimate the quality of extraction patterns. Further, we propose multiple solutions to control for semantic drift by taking advantage of the polarity of the categories to be learned (e.g., praiseworthy vs. blameworthy). Experimental results demonstrate that our algorithm achieves considerably better performance than several baselines.

Sentiment-Stance-Specificity (SSS) Dataset: Identifying Support-based Entailment among Opinions.

Pavithra Rajendran, Danushka Bollegala and Simon Parsons

Computational argumentation aims to model arguments as a set of premises that either support each other or collectively support a conclusion. We prepare three datasets of text-hypothesis pairs with support-based entailment based on opinions present in hotel reviews using a distant supervision approach. Support-based entailment is defined as the existence of a specific opinion (premise) that supports as well as entails a more general opinion and where these together support a generalised conclusion. A set of rules is proposed based on three different components — sentiment, stance and specificity to automatically predict support-based entailment. Two annotators manually annotated the relations among text-hypothesis pairs with an inter-rater agreement of 0.80. We compare the performance of the rules which gave an overall accuracy of 0.83. Further, we compare the performance of textual entailment under various conditions. The overall accuracy was 89.54%, 90.00% and 96.19% for our three datasets.

Resource Creation Towards Automated Sentiment Analysis in Telugu (a low resource language) and Integrating Multiple Domain Sources to Enhance Sentiment Prediction

Rama Rohit Reddy Gangula and Radhika Mamidi

Understanding the polarity or sentiment of a text is an important task in many application scenarios. Sentiment Analysis of a text can be used to answer various questions such as election prediction, favouredness towards any product etc. But the sentiment analysis task becomes challenging when it comes to low resource languages because the basis of learning sentiment classifiers are annotated datasets and annotated datasets for non-English texts hardly exists. So for the development of sentiment classifiers in Telugu, we have created corpora "Sentiraama" for different domains like movie reviews, song lyrics, product reviews and book reviews in Telugu language with the text written in

Telugu script. In this paper, we describe the process of creating the corpora and assigning polarities to them. After the creation of corpora, we trained the classifiers that yields good classification results. Typically a sentiment classifier is trained using data from the same domain it is intended to be tested on. But there may not be sufficient data available in the same domain and additionally using data from multiple sources and domains may help in creating a more generalized sentiment classifier which can be applied to multiple domains. So to create this generalized classifier, we used the sentiment data from the above corpus from different domains. We first tested the performance of sentiment analysis models built using single data source for both in-domain and cross-domain classification. Later, we built sentiment model using data samples from multiple domains and then tested the performance of the models based on their classification. Finally, we compared all the three approaches based on the performance of the models and discussed the best approach for sentiment analysis.

Multilingual Multi-class Sentiment Classification Using Convolutional Neural Networks

Mohammed Attia, Younes Samih, Ali Elkahky and Laura Kallmeyer

This paper describes a language-independent model for multi-class sentiment analysis using a simple neural network architecture of five layers (Embedding, Conv1D, GlobalMaxPooling and two Fully-Connected). The advantage of the proposed model is that it does not rely on language-specific features such as ontologies, dictionaries, or morphological or syntactic pre-processing. Equally important, our system does not use pre-trained word2vec embeddings which can be costly to obtain and train for some languages. In this research, we also demonstrate that oversampling can be an effective approach for correcting class imbalance in the data. We evaluate our methods on three publicly available datasets for English, German and Arabic, and the results show that our system's performance is comparable to, or even better than, the state of the art for these datasets. We make our source-code publicly available.

A Large Self-Annotated Corpus for Sarcasm

Mikhail Khodak, Nikunj Saunshi and Kiran Vodrahalli

We introduce the Self-Annotated Reddit Corpus (SARC), a large corpus for sarcasm research and for training and evaluating systems for sarcasm detection. The corpus has 1.3 million sarcastic statements — 10 times more than any previous dataset — and many times more instances of non-sarcastic statements, allowing for learning in both balanced and unbalanced label regimes. Each statement is furthermore self-annotated — sarcasm is labeled by the author, not an independent annotator — and provided with user, topic, and conversation context. We evaluate the corpus for accuracy, construct benchmarks for sarcasm detection, and evaluate baseline methods.

HappyDB: A Corpus of 100,000 Crowdsourced Happy Moments

Akari Asai, Sara Evensen, Behzad Golshan, Alon Halevy, Vivian Li, Andrei Lopatenko, Daniela Stepanov, Yoshihiko Suhara, Wang-Chiew Tan and Yinzhan Xu

The science of happiness is an area of positive psychology concerned with understanding what behaviors make people happy in a sustainable fashion. Recently, there has been interest in developing technologies that help incorporate the findings of the science of happiness into users' daily lives by steering them towards behaviors that increase happiness. With the goal of building technology that can understand how people express their happy moments in text, we crowd-sourced HappyDB, a corpus of 100,000 happy moments that we make publicly available. This paper describes HappyDB and its properties, and outlines several important NLP problems that can be studied with the help of the corpus. We also apply several state-of-the-art analysis techniques to analyze HappyDB. Our results demonstrate the need for deeper NLP techniques to be developed which makes HappyDB an exciting resource for follow-on research.

MultiBooked: A Corpus of Basque and Catalan Hotel Reviews Annotated for Aspect-level Sentiment Classification

Jeremy Barnes, Toni Badia and Patrik Lambert

While sentiment analysis has become an established field in the NLP community, research into languages other than English has been hindered by the lack of resources. Although much research in multi-lingual and cross-lingual sentiment analysis has focused on unsupervised or semi-supervised approaches, these still require a large number of resources and do not reach the performance of supervised approaches. With this in mind, we introduce two datasets for supervised aspect-level sentiment analysis in Basque and Catalan, both of which are under-resourced languages. We provide high-quality annotations and benchmarks with the hope that they will be useful to the growing community of researchers working on these languages.

Session P7 - Social Media Processing (1)

9th May 2018, 11:35

Chair person: **Paul Cook** Poster Session

BlogSet-BR: A Brazilian Portuguese Blog Corpus

Henrique Santos, Vinicius Woloszyn and Renata Vieira

The rich user-generated contend found on internet bloggs have always attracted the interest of scientific communities for many different purposes, such as from opinion and sentiment mining, information extraction or topic discovery. Nonetheless, a extensive corpora is essential to perform most of Natural Language Processing involved in these tasks. This paper presents BlogSet-BR, a extensive Brazilian Portuguese corpus containing 2.1 billions words extracted from 7.4 millions posts over 808 thousand different Brazilian blogs. Additionally, a extensible survey was conducted with authors to draw a profile of Brazilian bloggers.

SoMeWeTa: A Part-of-Speech Tagger for German Social Media and Web Texts

Thomas Proisl

Off-the-shelf part-of-speech taggers typically perform relatively poorly on web and social media texts since those domains are quite different from the newspaper articles on which most tagger models are trained. In this paper, we describe SoMeWeTa, a part-of-speech tagger based on the averaged structured perceptron that is capable of domain adaptation and that can use various external resources. We train the tagger on the German web and social media data of the EmpiriST 2015 shared task. Using the TIGER corpus as background data and adding external information about word classes and Brown clusters, we substantially improve on the state of the art for both the web and the social media data sets. The tagger is available as free software.

Collecting Code-Switched Data from Social Media

Gideon Mendels, Victor Soto, Aaron Jaech and Julia Hirschberg

We address the problem of mining code-switched data from the web, where code-switching is defined as the tendency of bilinguals to switch between their multiple languages both across and within utterances. We propose a method that identifies data as code-switched in languages L1 and L2 when a language classifier labels the document as language L1 but the document also contains words that can only belong to L2. We apply our method to Twitter data and collect a set of more than 43,000 tweets. We obtain language identifiers for a subset of 8,000 tweets using crowd-sourcing with high inter-annotator agreement and accuracy. We validate our Twitter corpus by comparing it to the Spanish-English corpus of code-switched tweets collected for the EMNLP 2016 Shared Task for Language Identification, in terms of code-switching rates, language composition and amount of code-switch types found in both datasets. We then trained language taggers on both corpora and show that a tagger trained on the EMNLP corpus exhibits a considerable drop in accuracy when tested on the new corpus and a tagger trained on our new corpus achieves very high accuracy when tested on both corpora.

Classifying the Informative Behaviour of Emoji in Microblogs

Giulia Donato and Patrizia Paggio

Emoji are pictographs commonly used in microblogs as emotion markers, but they can also represent a much wider range of concepts. Additionally, they may occur in different positions within a message (e.g. a tweet), appear in sequences or act as word substitute. Emoji must be considered necessary elements in the analysis and processing of user generated content, since they can either provide fundamental syntactic information, emphasize what is already expressed in the text, or carry meaning that cannot be inferred from the words alone. We collected and annotated a corpus of 2475 tweets pairs with the aim of analyzing and then classifying emoji use with respect to redundancy. The best classification model achieved an F-score of 0.7. In this paper we shortly present the corpus, and we describe the classification experiments, explain the predictive features adopted, discuss the problematic aspects of our approach and suggest future improvements.

A Taxonomy for In-depth Evaluation of Normalization for User Generated Content

Rob Van der Goot, Rik Van Noord and Gertjan Van Noord

In this work we present a taxonomy of error categories for lexical normalization, which is the task of translating user generated content to canonical language. We annotate a recent normalization dataset to test the practical use of the taxonomy and read a near-perfect agreement. This annotated dataset is then used to evaluate how an existing normalization model performs on the different categories of the taxonomy. The results of this evaluation reveal that some of the problematic categories only include minor transformations, whereas most regular transformations are solved quite well.

Gaining and Losing Influence in Online Conversation

Arun Sharma and Tomek Strzalkowski

In this paper, we describe a study we conducted to determine, if a person who is highly influential in a discussion on a familiar topic would retain influence when moving to a topic that is less familiar or perhaps not as interesting. For this research, we collected samples of realistic on-line chat room discussions on several topics related to current issues in education, technology, arts, sports, finances, current affairs, etc. The collected data allowed us to create models for specific types of conversational behavior, such as agreement, disagreement, support, persuasion, negotiation, etc. These models were used to study influence in online discussions. It also allowed us to study how human influence works in online discussion and what affects a person's influence from one topic to another. We found that influence is impacted by topic familiarity, sometimes dramatically, and we explain how it is affected and why.

Arap-Tweet: A Large Multi-Dialect Twitter Corpus for Gender, Age and Language Variety Identification

Wajdi Zaghouani and Anis Charfi

In this paper, we present Arap-Tweet, which is a large-scale and multi-dialectal corpus of Tweets from 11 regions and 16 countries in the Arab world representing the major Arabic dialectal varieties. To build this corpus, we collected data from Twitter and we provided a team of experienced annotators with annotation guidelines that they used to annotate the corpus for age categories, gender, and dialectal variety. During the data collection effort, we based our search on distinctive keywords that are specific to the different Arabic dialects and we also validated the location using Twitter API. In this paper, we report on the corpus data collection and annotation efforts. We also present some issues that we encountered during these phases. Then, we present the results of the evaluation performed to ensure the consistency of the annotation. The provided corpus will enrich the limited set of available language resources for Arabic and will be an invaluable enabler for developing author profiling tools and NLP tools for Arabic.

Session O5 - Language Resource Policies & Management

9th May 2018, 14:35

Chair person: **Stelios Piperidis** Oral Session

Data Management Plan (DMP) for Language Data under the New General Data Protection Regulation (GDPR)

Pawel Kamocki, Valérie Mapelli and Khalid Choukri

The EU's General Data Protection Regulation (GDPR) of 27 April 2016 will apply from 25 May 2018. It will reinforce certain principles related to the processing of personal data, which will also affect many projects in the field of Natural Language Processing. Perhaps most importantly, the GDPR will introduce the principle of accountability, according to which the data processor shall be able to demonstrate compliance with the new rules, and that he applies 'privacy by design and by default'. In our opinion, a well-drafted Data Management Plan (DMP) is of key importance for GDPR compliance; indeed, the trend towards the adoption of a DMP, particularly in EU-funded research projects, has been more vivid since 2017, after the extension of the Horizon 2020 Open Data Pilot. Since 2015, ELRA also proposes its own template for the Data Management Plan, which is being updated to take the new law into account. In this paper, we present the new legal framework introduced by the GDPR and propose how the new rules can be integrated in the DMP in order to increase transparency of processing, facilitate demonstration of GDPR compliance and spread good practices within the community.

We Are Depleting Our Research Subject as We Are Investigating It: In Language Technology, more Replication and Diversity Are Needed

António Branco

In this paper, we present an analysis indicating that, in language technology, as we are investigating natural language we are contributing to deplete it in the sense that we are contributing to reduce the diversity of languages. To address this circumstance, we propose that more replication and reproduction and more language diversity need to be taken into account in our research activities.

Lessons Learned: On the Challenges of Migrating a Research Data Repository from a Research Institution to a University Library.

Thorsten Trippel and Claus Zinn

The transfer of research data management from one institution to another infrastructural partner is all but trivial, but can be

required, for instance, when an institution faces reorganisation or closure. In a case study, we describe the migration of all research data, identify the challenges we encountered, and discuss how we addressed them. It shows that the moving of research data management to another institution is a feasible, but potentially costly enterprise. Being able to demonstrate the feasibility of research data migration supports the stance of data archives that users can expect high levels of trust and reliability when it comes to data safety and sustainability.

Introducing NIEUW: Novel Incentives and Workflows for Eliciting Linguistic Data

Christopher Cieri, James Fiumara, Mark Liberman, Chris Callison-Burch and Jonathan Wright

This paper introduces the NIEUW (Novel Incentives and Workflows) project funded by the United States National Science Foundation and part of the Linguistic Data Consortium's strategy to provide order of magnitude improvement in the scale, cost, variety, linguistic diversity and quality of Language Resources available for education, research and technology development. Notwithstanding decades of effort and progress in collecting and distributing Language Resources, it remains the case that demand still far exceeds supply for all of the approximately 7000 languages in the world, even the most well documented languages with global economic and political influence. The absence of Language Resources, regardless of the language, stifles teaching and technology building, inhibiting the creation of language enabled applications and, as a result, commerce and communication. Project oriented approaches which focus intensive funding and effort on problems of limited scope over short durations can only address part of the problem. The HLT community instead requires approaches that do not rely upon highly constrained resources such as project funding and can be sustained across many languages and many years. In this paper, we describe a new initiative to harness the power of alternative incentives to elicit linguistic data and annotation. We also describe changes to the workflows necessary to collect data from workforces attracted by these incentives.

Three Dimensions of Reproducibility in Natural Language Processing

K. Bretonnel Cohen, Jingbo Xia, Pierre Zweigenbaum, Tiffany Callahan, Orin Hargraves, Foster Goss, Nancy Ide, Aurélie Névéol, Cyril Grouin and Lawrence E. Hunter

Despite considerable recent attention to problems with reproducibility of scientific research, there is a striking lack of agreement about the definition of the term. That is a problem, because the lack of a consensus definition makes it difficult to compare studies of reproducibility, and thus to have even a broad overview of the state of the issue in natural language processing. This paper proposes an ontology of reproducibility in that field. Its goal is to enhance both future research and communication about the topic, and retrospective meta-analyses. We show that three dimensions of reproducibility, corresponding to three kinds of claims in natural language processing papers, can account for a variety of types of research reports. These dimensions are reproducibility of a conclusion, of a finding, and of a value. Three biomedical natural language processing papers by the authors of this paper are analyzed with respect to these dimensions.

Session O6 - Emotion & Sentiment (1)

9th May 2018, 14:35

Chair person: **Pushpak Bhattacharyya** Oral Session

Content-Based Conflict of Interest Detection on Wikipedia

Udochukwu Orizu and Yulan He

Wikipedia is one of the most visited websites in the world. On Wikipedia, Conflict-of-Interest (CoI) editing happens when an editor uses Wikipedia to advance their interests or relationships. This includes paid editing done by organisations for public relations purposes, etc. CoI detection is highly subjective and though closely related to vandalism and bias detection, it is a more difficult problem. In this paper, we frame CoI detection as a binary classification problem and explore various features which can be used to train supervised classifiers for CoI detection on Wikipedia articles. Our experimental results show that the best F-measure achieved is 0.67 by training SVM from a combination of features including stylometric, bias and emotion features. As we are not certain that our non-CoI set does not contain any CoI articles, we have also explored the use of one-class classification for CoI detection. The results show that using stylometric features outperforms other types of features or a combination of them and gives an F-measure of 0.63. Also, while binary classifiers give higher recall values (0.81~0.94), one-class classifier attains higher precision values (0.69~0.74).

Word Affect Intensities

Saif Mohammad

Words often convey affect—emotions, feelings, and attitudes. Further, different words can convey affect to various degrees (intensities). However, existing manually created lexicons for basic emotions (such as anger and fear) indicate only coarse categories of affect association (for example, associated with anger or not associated with anger). Automatic lexicons of affect provide fine degrees of association, but they tend not to be accurate

as human-created lexicons. Here, for the first time, we present a manually created affect intensity lexicon with real-valued scores of intensity for four basic emotions: anger, fear, joy, and sadness. (We will subsequently add entries for more emotions such as disgust, anticipation, trust, and surprise.) We refer to this dataset as the {\it NRC Affect Intensity Lexicon}, or {\it AIL} for short. AIL has entries for close to 6,000 English words. We used a technique called best–worst scaling (BWS) to create the lexicon. BWS improves annotation consistency and obtains reliable fine-grained scores (split-half reliability > 0.91). We also compare the entries in AIL with the entries in the {\it NRC VAD Lexicon}, which has valence, arousal, and dominance (VAD) scores for 20K English words. We find that anger, fear, and sadness words, on average, have very similar VAD scores. However, sadness words tend to have slightly lower dominance scores than fear and anger words. The Affect Intensity Lexicon has applications in automatic emotion analysis in a number of domains such as commerce, education, intelligence, and public health. AIL is also useful in the building of natural language generation systems.

Representation Mapping: A Novel Approach to Generate High-Quality Multi-Lingual Emotion Lexicons

Sven Buechel and Udo Hahn

In the past years, sentiment analysis has increasingly shifted attention to representational frameworks more expressive than semantic polarity (being positive, negative or neutral). However, these richer formats (like Basic Emotions or Valence-Arousal-Dominance, and variants therefrom), rooted in psychological research, tend to proliferate the number of representation schemes for emotion encoding. Thus, a large amount of representationally incompatible emotion lexicons has been developed by various research groups adopting one or the other emotion representation format. As a consequence, the reusability of these resources decreases as does the comparability of systems using them. In this paper, we propose to solve this dilemma by methods and tools which map different representation formats onto each other for the sake of mutual compatibility and interoperability of language resources. We present the first large-scale investigation of such representation mappings for four typologically diverse languages and find evidence that our approach produces (near-)gold quality emotion lexicons, even in crosslingual settings. Finally, we use our models to create new lexicons for eight typologically diverse languages.

Unfolding the External Behavior and Inner Affective State of Teammates through Ensemble Learning: Experimental Evidence from a Dyadic Team Corpus

Aggeliki Vlachostergiou, Mark Dennison, Catherine Neubauer, Stefan Scherer, Peter Khooshabeh and Andre Harrison

The current study was motivated to understand the relationship between the external behavior and inner affective state of two team members (“instructor"-“defuser") during a demanding operational task (i.e., bomb defusion). In this study we assessed team member’s verbal responses (i.e., length of duration) in relation to their external as well as internal affective cues. External behavioral cues include defuser’s verbal expressions while inner cues are based on physiological signals. More specifically, we differentiate between “defusers’" physiological patterns occurring after the “instructor’s" turns according to whether they belong to a short or a long turn-taking response interval. Based on the assumption that longer turn-taking behaviors are likely to be caused by demanding cognitive task events and/or stressful interactions, we hypothesize that inner mechanisms produced in these intense affective activity intervals will be reflected on defuser’s physiology. A dyadic team corpus was used to examine the association between the “defusers" physiological signals following the “instructor’s" questions to predict whether they occurred in a short or long turn-taking period of time. The results suggest that an association does exist between turn taking and inner affective state. Additionally, it was our goal to further unpack this association by creating diverse ensembles. As such, we studied various base learners and different ensemble sizes to determine the best approach towards building a stable diverse ensemble that generalizes well on the external and inner cues of individuals.

Understanding Emotions: A Dataset of Tweets to Study Interactions between Affect Categories

Saif Mohammad and Svetlana Kiritchenko

Human emotions are complex and nuanced. Yet, an overwhelming majority of the work in automatically detecting emotions from text has focused only on classifying text into positive, negative, and neutral classes. Our goal is to create a single textual dataset that is annotated for many emotion dimensions (from both the basic emotion model and the VAD model). For each emotion dimension, we annotate tweets for not just coarse classes (such as anger or no anger) but also for fine-grained real-valued scores indicating the intensity of emotion. We use Best-Worst Scaling to address the limitations of traditional rating scale methods such as inter- and intra-annotator inconsistency. We

show that the fine-grained intensity scores thus obtained are reliable. The new dataset is useful for training and testing supervised machine learning algorithms for multi-label emotion classification, emotion intensity regression, detecting valence, detecting ordinal class of intensity of emotion (slightly sad, very angry, etc.), and detecting ordinal class of valence. The dataset also sheds light on crucial research questions such as: which emotions often present together in tweets?; how do the intensities of the three negative emotions relate to each other?; and how do the intensities of the basic emotions relate to valence?

Session O7 - Knowledge Discovery & Evaluation (1)

9th May 2018, 14:35

Chair person: **Andrejs Vasiljevs** Oral Session

When ACE met KBP: End-to-End Evaluation of Knowledge Base Population with Component-level Annotation

Bonan Min, Marjorie Freedman, Roger Bock and Ralph Weischedel

Building a Knowledge Base from text corpora is useful for many applications such as question answering and web search. Since 2012, the Cold Start Knowledge Base Population (KBP) evaluation at the Text Analysis Conference (TAC) has attracted many participants. Despite the popularity, the Cold Start KBP evaluation has several problems including but not limited to the following two: first, each year's assessment dataset is a pooled set of query-answer pairs, primarily generated by participating systems. It is well known to participants that there is pooling bias: a system developed outside of the official evaluation period is not rewarded for finding novel answers, but rather is penalized for doing so. Second, the assessment dataset, constructed with lots of human effort, offers little help in training information extraction algorithms which are crucial ingredients for the end-to-end KBP task. To address these problems, we propose a new unbiased evaluation methodology that uses existing component-level annotation such as the Automatic Content Extraction (ACE) dataset, to evaluate Cold Start KBP. We also propose bootstrap resampling to provide statistical significance to the results reported. We will then present experimental results and analysis.

Simple Large-scale Relation Extraction from Unstructured Text

Christos Christodoulopoulos and Arpit Mittal

Knowledge-based question answering relies on the availability of facts, the majority of which cannot be found in structured sources (e.g. Wikipedia info-boxes, Wikidata). One of the major components of extracting facts from unstructured text is Relation Extraction (RE). In this paper we propose a novel method for creating distant (weak) supervision labels for training a large-scale RE system. We also provide new evidence about the effectiveness of neural network approaches by decoupling the model architecture from the feature design of a state-of-the-art neural network system. Surprisingly, a much simpler classifier trained on similar features performs on par with the highly complex neural network system (at 75x reduction to the training time), suggesting that the features are a bigger contributor to the final performance.

Joint Learning of Sense and Word Embeddings

Mohammed Alsuhaibani and Danushka Bollegala

Methods for learning lower-dimensional representations (embeddings) of words using unlabelled data have received a renewed interested due to their myriad success in various Natural Language Processing (NLP) tasks. However, despite their success, a common deficiency associated with most word embedding learning methods is that they learn a single representation for a word, ignoring the different senses of that word (polysemy). To address the polysemy problem, we propose a method that jointly learns sense-aware word embeddings using both unlabelled and sense-tagged text corpora. In particular, our proposed method can learn both word and sense embeddings by efficiently exploiting both types of resources. Our quantitative and qualitative experimental results using unlabelled text corpus with (a) manually annotated word senses, and (b) pseudo annotated senses demonstrate that the proposed method can correctly learn the multiple senses of an ambiguous word. Moreover, the word embeddings learnt by our proposed method outperform several previously proposed competitive word embedding learning methods on word similarity and short-text classification benchmark datasets.

Comparing Pretrained Multilingual Word Embeddings on an Ontology Alignment Task

Dagmar Gromann and Thierry Declerck

Word embeddings capture a string's semantics and go beyond its surface form. In a multilingual environment, those embeddings need to be trained for each language, either separately or as a joint model. The more languages needed, the more computationally cost- and time-intensive the task of training. As an alternative, pretrained word embeddings can be utilized to compute semantic similarities of strings in different languages. This paper provides a comparison of three different multilingual pretrained word embedding repositories with a string-matching baseline and uses the task of ontology alignment as example scenario. A vast majority of ontology alignment methods rely on string similarity

metrics, however, they frequently use string matching techniques that purely rely on syntactic aspects. Semantically oriented word embeddings have much to offer to ontology alignment algorithms, such as the simple Munkres algorithm utilized in this paper. The proposed approach produces a number of correct alignments on a non-standard data set based on embeddings from the three repositories, where FastText embeddings performed best on all four languages and clearly outperformed the string-matching baseline.

A Large Resource of Patterns for Verbal Paraphrases

Octavian Popescu, Ngoc Phuoc An Vo and Vadim Sheinin

Paraphrases play an important role in natural language understanding, especially because there are fluent jumps between hidden paraphrases in a text. For example, even to get to the meaning of a simple dialog like I bought a computer. How much did the computer cost? involves quite a few steps. A computational system may actually have a huge problem in linking the two sentences as their connection is not overtly present in the text. However, it becomes easier if it has access to the following paraphrases: [HUMAN] buy [ARTIFACT] () [HUMAN] pay [PRICE] for [ARTIFACT] () [ARTIFACT] cost [HUMAN] [PRICE], and also to the information that I IsA [HUMAN] and computer IsA [ARTIFACT]. In this paper we introduce a resource of such paraphrases that was extracted by investigating large corpora in an unsupervised manner. The resource contains tens of thousands of such pairs and it is available for academic purposes.

Session O8 - Corpus Creation, Use & Evaluation (1)

9th May 2018, 14:35

Chair person: **Patrizia Paggio** Oral Session

Building Parallel Monolingual Gan Chinese Dialects Corpus

Fan Xu, Mingwen Wang and Maoxi Li

Automatic language identification of an input sentence or a text written in similar languages, varieties or dialects is an important task in natural language processing. In this paper, we propose a scheme to represent Gan (Jiangxi province of China) Chinese dialects. In particular, it is a two-level and fine-grained representation using Chinese character, Chinese Pinyin and Chinese audio forms. Guided by the scheme, we manually annotate a Gan Chinese Dialects Corpus (GCDC) including 131.5 hours and 310 documents with 6 different genres, containing news, official document, story, prose, poet, letter and speech, from 19 different Gan regions. In addition, the preliminary evaluation on 2-way, 7-way and 20-way sentence-level Gan Chinese Dialects Identification (GCDI) justifies the appropriateness of the scheme to Gan Chinese dialects analysis and the usefulness of our manually annotated GCDC.

A Recorded Debating Dataset

Shachar Mirkin, Michal Jacovi, Tamar Lavee, Hong-Kwang Kuo, Samuel Thomas, Leslie Sager, Lili Kotlerman, Elad Venezian and Noam Slonim

This paper describes an English audio and textual dataset of debating speeches, a unique resource for the growing research field of computational argumentation and debating technologies. We detail the process of speech recording by professional debaters, the transcription of the speeches with an Automatic Speech Recognition (ASR) system, their consequent automatic processing to produce a text that is more "NLP-friendly", and in parallel – the manual transcription of the speeches in order to produce gold-standard "reference" transcripts. We release 60 speeches on various controversial topics, each in five formats corresponding to the different stages in the production of the data. The intention is to allow utilizing this resource for multiple research purposes, be it the addition of in-domain training data for a debate-specific ASR system, or applying argumentation mining on either noisy or clean debate transcripts. We intend to make further releases of this data in the future.

Building a Corpus from Handwritten Picture Postcards: Transcription, Annotation and Part-of-Speech Tagging

Kyoko Sugisaki, Nicolas Wiedmer and Heiko Hausendorf

In this paper, we present a corpus of over 11,000 holiday picture postcards written in German and Swiss German. The postcards have been collected for the purpose of text-linguistic investigations on the genre and its standardisation and variation over time. We discuss the processes and challenges of digitalisation, manual transcription, and manual annotation. In addition, we developed our own automatic text segmentation system and a part-of-speech tagger, since our texts often contain orthographic deviations, domain-specific structures such as fragments, subject-less sentences, interjections, discourse particles, and domain-specific formulaic communicative routines in salutation and greeting. In particular, we demonstrate that the CRF-based POS tagger could be boosted to a domain-specific text by adding a small amount of in-domain data. We showed that entropy-based training data sampling was competitive with random sampling in performing this task. The evaluation showed that our POS tagger achieved a F1 score of 0.93 (precision 0.94, recall 0.93), which outperformed a state-of-the-art POS tagger.

A Lexical Tool for Academic Writing in Spanish based on Expert and Novice Corpora

Marcos García Salido, Marcos Garcia, Milka Villayandre-Llamazares and Margarita Alonso-Ramos

The object of this article is to describe the extraction of data from a corpus of academic texts in Spanish and the use of those data for developing a lexical tool oriented to the production of academic texts. The corpus provides the lexical combinations that will be included in the afore-mentioned tool, namely collocations, idioms and formulas. They have been retrieved from the corpus controlling for their keyness (i.e., their specificity with regard to academic texts) and their even distribution across the corpus. For the extraction of collocations containing academic vocabulary other methods have been used, taking advantage of the morphological and syntactic information with which the corpus has been enriched. In the case of collocations and other multiword units, several association measures are being tested in order to restrict the list of candidates the lexicographers will have to deal with manually.

Framing Named Entity Linking Error Types

Adrian Brasoveanu, Giuseppe Rizzo, Philipp Kuntschick, Albert Weichselbraun and Lyndon J.B. Nixon

Named Entity Linking (NEL) and relation extraction forms the backbone of Knowledge Base Population tasks. The recent rise oflarge open source Knowledge Bases and the continuous focus on improving NEL performance has led to the creation of automatedbenchmark solutions during the last decade. The benchmarking of NEL systems offers a valuable approach to understand a NELsystem's performance quantitatively. However, an in-depth qualitative analysis that helps improving NEL methods by identifying errorcauses usually requires a more thorough error analysis. This paper proposes a taxonomy to frame common errors and applies thistaxonomy in a survey study to assess the performance of four well-known Named Entity Linking systems on three recent gold standards.

Session P8 - Character Recognition and Annotation

9th May 2018, 14:35

Chair person: **Jordi Turmo** Poster Session

Transc&Anno: A Graphical Tool for the Transcription and On-the-Fly Annotation of Handwritten Documents

Nadezda Okinina, Lionel Nicolas and Verena Lyding

We present Transc&Anno, a web-based collaboration tool allowing the transcription of text images and their shallow on-the-fly annotation. Transc&Anno was originally developed in order to address the needs of learner corpora research so as to facilitate digitisation of handwritten learner essays. However, the tool can be used for the creation of any type of corpora requiring transcription and shallow on-the-fly annotation resulting in inline XML. Transc&Anno provides an intuitive environment that is explicitly designed to facilitate the transcription and annotation process for linguists. Transc&Anno ensures a high transcription output quality by validating the XML and only allowing predefined tags. It was created on top of the FromThePage transcription tool developed entirely with standard web technologies – Ruby on Rails, Javascript, HTML, and CSS. We adapted this open-source web-based tool to linguistic research purposes by adding linguistic annotation functionalities to it. Thereby we united the convenience of a collaborative transcription tool with its advanced image visualisation, centralised data storage, version control and inter-collaborator communication facilities with the precision of a linguistic annotation tool with its well-developed tag definition possibilities, easy tagging process and tagged-text visualisation. Transc&Anno is easily customisable, open source, and available on Github.

Correction of OCR Word Segmentation Errors in Articles from the ACL Collection through Neural Machine Translation Methods

Vivi Nastase and Julian Hitschler

Depending on the quality of the original document, Optical Character Recognition (OCR) can produce a range of errors – from erroneous letters to additional and spurious blank spaces. We applied a sequence-to-sequence machine translation system to correct word-segmentation OCR errors in scientific texts from the ACL collection with an estimated precision and recall above 0.95 on test data. We present the correction process and results.

From Manuscripts to Archetypes through Iterative Clustering

Armin Hoenen

Corpora of manuscripts of the same ancient text often preserve many variants. This is so because upon copying over long copy chains errors and editorial changes have been repeatedly made and reverted to the effect that most often no 2 variant texts of the same so-called textual tradition have exactly the same text. Obviously in order to save the time to read all of the versions and in order to enable discourse and unambiguous referencing, philologists have since the beginnings of the age of print embarked on providing one single textual representation of this variety. In computational terms one tries to retrieve/compose the base text which is most likely the latest common ancestor (archetype) of all observed variants. Computationally, stemmata – that is trees depicting

the copy history (manuscripts = nodes, Copy processes = edges) – have been computed and evaluated automatically (Roos and Heikkilä, 2009). Likewise, automatic archetype reconstruction has been introduced lately, (Hoenen, 2015b; Koppel et al., 2016). A synthesis of both stemma generation and archetype reconstruction has not yet been achieved. This paper therefore presents an approach where through iterative clustering a stemma and an archetype text are being reconstructed bottom-up.

Building A Handwritten Cuneiform Character Imageset

Kenji Yamauchi, Hajime Yamamoto and Wakaha Mori

Digitization of cuneiform documents is important to boost the research activity on ancient Middle East and some projects have been launched in around 2,000. However, the digitization process is laborious due to the huge scale of the documents and no trustful (semi-)automatic method has established. In this paper, we focused on a cuneiform document digitization task, realization of Optical Character Recognition (OCR) method from the handwritten copies of original materials. Currently, as the first step toward development of such methods, we are constructing a handwritten cuneiform character imageset named with professional assistance. This imageset contains typical stroke patterns for handwriting each frequently appearing cuneiform character and will be able to support the development of handwritten cuneiform OCR system.

PDF-to-Text Reanalysis for Linguistic Data Mining

Michael Wayne Goodman, Ryan Georgi and Fei Xia

Extracting semi-structured text from scientific writing in PDF files is a difficult task that has faced researchers for decades. In the 1990s, this task was largely a computer vision and OCR problem, as PDF files were often the result of scanning printed documents. Today, PDFs have standardized digital typesetting without the need for OCR, but extraction of semi-structured text from these documents remains a nontrivial task. In this paper, we present a system for the reanalysis of glyph-level PDF extracted text that performs block detection, respacing, and tabular data analysis for the purposes of linguistic data mining. We further present our reanalyzed output format, which attempts to eliminate the extreme verbosity of XML output while leaving important positional information available for downstream processes.

Session P9 - Conversational Systems/Dialogue/Chatbots/Human-Robot Interaction (1)

9th May 2018, 14:35

Chair person: **Leo Wanner** Poster Session

Crowdsourced Multimodal Corpora Collection Tool

Patrik Jonell, Catharine Oertel, Dimosthenis Kontogiorgos, Jonas Beskow and Joakim Gustafson

In recent years, more and more multimodal corpora have been created. To our knowledge there is no publicly available tool which allows for acquiring controlled multimodal data of people in a rapid and scalable fashion. We therefore are proposing (1) a novel tool which will enable researchers to rapidly gather large amounts of multimodal data spanning a wide demographic range, and (2) an example of how we used this tool for corpus collection of our “Attentive listener” multimodal corpus. The code is released under an Apache License 2.0 and available as an open-source repository, which can be found at \url{https://github.com/kth-social-robotics/multimodal-crowdsourcing-tool}. This tool will allow researchers to set-up their own multimodal data collection system quickly and create their own multimodal corpora. Finally, this paper provides a discussion about the advantages and disadvantages with a crowd-sourced data collection tool, especially in comparison to a lab recorded corpora.

Expert Evaluation of a Spoken Dialogue System in a Clinical Operating Room

Juliana Miehle, Nadine Gerstenlauer, Daniel Ostler, Hubertus Feußner, Wolfgang Minker and Stefan Ultes

With the emergence of new technologies, the surgical working environment becomes increasingly complex and comprises many medical devices which have to be monitored and controlled. With the aim of improving productivity and reducing the workload for the operating staff, we have developed an Intelligent Digital Assistant for Clinical Operating Rooms (IDACO) which allows the surgeon to control the operating room using natural spoken language. As speech is the modality used by the surgeon to communicate with their staff, using it to control the technical devices does not pose an additional mental burden. Therefore, we claim that the surgical environment presents a potential field of application for Spoken Dialogue Systems. In this work, we present the design and implementation of IDACO as well as the evaluation in an experimental set-up by specialists in the field of minimally invasive surgery. Our expert evaluation yields

promising results and allows to conclude that clinical operating rooms are indeed an expedient area of application for Spoken Dialogue Systems.

JAIST Annotated Corpus of Free Conversation

Kiyoaki Shirai and Tomotaka Fukuoka

This paper introduces an annotated corpus of free conversations in Japanese. It is manually annotated with two kinds of linguistic information: dialog act and sympathy. First, each utterance in the free conversation is annotated with its dialog act, which is chosen from a coarse-grained set consisting of nine dialog act labels. Cohen's kappa of the dialog act annotation between two annotators was 0.636. Second, each utterance is judged whether the speaker expresses his/her sympathy or antipathy toward the other participant or the current topic in the conversation. Cohen's kappa of sympathy tagging was 0.27, indicating the difficulty of the sympathy identification task. As a result, the corpus consists of 92,031 utterances in 97 dialogs. Our corpus is the first annotated corpus of Japanese free conversations that is publicly available.

The Metalogue Debate Trainee Corpus: Data Collection and Annotations

Volha Petukhova, Andrei Malchanau, Youssef Oualil, Dietrich Klakow, Saturnino Luz, Fasih Haider, Nick Campbell, Dimitris Koryzis, Dimitris Spiliotopoulos, Pierre Albert, Nicklas Linz and Jan Alexandersson

This paper describes the Metalogue Debate Trainee Corpus (DTC). DTC has been collected and annotated in order to facilitate the design of instructional and interactive models for Virtual Debate Coach application - an intelligent tutoring system used by young parliamentarians to train their debate skills. The training is concerned with the use of appropriate multimodal rhetorical devices in order to improve (1) the organization of arguments, (2) arguments' content selection, and (3) argument delivery techniques. DTC contains tracking data from motion and speech capturing devices and semantic annotations - dialogue acts - as defined in ISO 24617-2 and discourse relations as defined in ISO 24617-8. The corpus comes with a manual describing the data collection process, annotation activities including an overview of basic concepts and their definitions including annotation schemes and guidelines on how to apply them, tools and other resources. DTC will be released in the ELRA catalogue in second half of 2018.

Towards Continuous Dialogue Corpus Creation: writing to corpus and generating from it

Andrei Malchanau, Volha Petukhova and Harry Bunt

This paper describes a method to create dialogue corpora annotated with interoperable semantic information. The corpus development is performed following the ISO linguistic annotation framework and primary data encoding initiatives. The Continuous Dialogue Corpus Creation (D3C) methodology is proposed, where a corpus is used as a shared repository for analysis and modelling of interactive dialogue behaviour, and for implementation, integration and evaluation of dialogue system components. All these activities are supported by the use of ISO standard data models including annotation schemes, encoding formats, tools, and architectures. Standards also facilitate practical work in dialogue system implementation, deployment, evaluation and re-training, and enabling automatic generation of adequate system behaviour from the data. The proposed methodology is applied to the data-driven design of two multimodal interactive applications - the Virtual Negotiation Coach, used for the training of metacognitive skills in a multi-issue bargaining setting, and the Virtual Debate Coach, used for the training of debate skills in political contexts.

MYCanCor: A Video Corpus of spoken Malaysian Cantonese

Andreas Liesenfeld

The Malaysia Cantonese Corpus (MYCanCor) is a collection of recordings of Malaysian Cantonese speech mainly collected in Perak, Malaysia. The corpus consists of around 20 hours of video recordings of spontaneous talk-in-interaction (56 settings) typically involving 2-4 speakers. A short scene description as well as basic speaker information is provided for each recording. The corpus is transcribed in CHAT (minCHAT) format and presented in traditional Chinese characters (UTF8) using the Hong Kong Supplementary Character Set (HKSCS). MYCanCor is expected to be a useful resource for researchers interested in any aspect of spoken language processing or Chinese multimodal corpora.

KTH Tangrams: A Dataset for Research on Alignment and Conceptual Pacts in Task-Oriented Dialogue

Todd Shore, Theofronia Androulakaki and Gabriel Skantze

There is a growing body of research focused on task-oriented instructor-manipulator dialogue, whereby one dialogue participant initiates a reference to an entity in a common environment while the other participant must resolve this reference in order to manipulate said entity. Many of these works are based on disparate if nevertheless similar datasets. This paper described an English corpus of referring expressions in relatively free, unrestricted dialogue with physical features generated in a simulation, which facilitate analysis of dialogic linguistic phenomena regarding alignment in the formation of referring expressions known as conceptual pacts.

On the Vector Representation of Utterances in Dialogue Context

Louisa Pragst, Niklas Rach, Wolfgang Minker and Stefan Ultes

In recent years, the representation of words as vectors in a vector space, also known as word embeddings, has achieved a high degree of attention in the research community and the benefits of such a representation can be seen in the numerous applications that utilise it. In this work, we introduce dialogue vector models, a new language resource that represents dialogue utterances in vector space and captures the semantic meaning of those utterances in the dialogue context. We examine how the word vector approach can be applied to utterances in a dialogue to generate a meaningful representation of them in vector space. Utilising existing dialogue corpora and word vector models, we create dialogue vector models and show that they capture relevant semantic information by comparing them to manually annotated dialogue acts. Furthermore, we discuss potential areas of application for dialogue vector models, such as dialogue act annotation, learning of dialogue strategies, intent detection and paraphrasing.

ES-Port: a Spontaneous Spoken Human-Human Technical Support Corpus for Dialogue Research in Spanish

Laura García-Sardiña, Manex Serras and Arantza Del Pozo

In this paper the ES-Port corpus is presented. ES-Port is a spontaneous spoken human-human dialogue corpus in Spanish that consists of 1170 dialogues from calls to the technical support department of a telecommunications provider. This paper describes its compilation process, from the transcription of the raw audio to the anonymisation of the sensitive data contained in the transcriptions. Because the anonymisation process was carried out through substitution by entities of the same type, coherence and readability are kept within the anonymised dialogues. In the resulting corpus, the replacements of the anonymised entities are labelled with their corresponding categories. In addition, the corpus is annotated with acoustic-related extralinguistic events such as background noise or laughter and linguistic phenomena such as false starts, use of filler words or code switching. The ES-Port corpus is now publicly available through the META-SHARE repository, with the main objective of promoting further research into more open domain data-driven dialogue systems in Spanish.

From analysis to modeling of engagement as sequences of multimodal behaviors

Soumia Dermouche and Catherine Pelachaud

In this paper, we present an approach to endow an Embodied Conversational Agent with engagement capabilities. We relied on a corpus of expert-novice interactions. Two types of manual annotation were conducted: non-verbal signals such as gestures, head movements and smiles; engagement level of both expert and novice during the interaction. Then, we used a temporal sequence mining algorithm to extract non-verbal sequences eliciting variation of engagement perception. Our aim is to apply these findings in human-agent interaction to analyze user's engagement level and to control agent's behavior. The novelty of this study is to consider explicitly engagement as sequence of multimodal behaviors.

Session P10 - Digital Humanities

9th May 2018, 14:35

Chair person: **Giorgio Maria Di Nunzio** Poster Session

A corpus of German political speeches from the 21st century

Adrien Barbaresi

The present German political speeches corpus follows from a initial release which has been used in various research contexts. This article documents an updated and extended version: as 2017 marks the end of a legislative period, the corpus now includes the four highest ranked functions on federal state level. Besides providing a citable reference for this resource, the main contributions are (1) an extensive description of the corpus to be released and (2) the description of an interface to navigate through the texts, designed for researchers beyond the corpus and computational linguistics communities as well as for the general public. The corpus can be considered to be from the 21st century since most speeches have been written after 2001 and also because it includes a visualization interface providing synoptic overviews ordered chronologically, by speaker or by keyword as well as consequent accesses to the texts.

Building Literary Corpora for Computational Literary Analysis - A Prototype to Bridge the Gap between CL and DH

Andrew Frank and Christine IVANOVIC

The design of LitText follows the traditional research approach in digital humanities (DH): collecting texts for critical reading and underlining parts of interest. Texts, in multiple languages, are prepared with a minimal markup language, and processed by NLP

services. The result is converted to RDF (a.k.a. semantic-web, linked-data) triples. Additional data available as linked data on the web (e.g. Wikipedia data) can be added. The DH researcher can then harvest the corpus with SPARQL queries. The approach is demonstrated with the construction of a 20 million word corpus from English, German, Spanish, French and Italian texts and an example query to identify texts where animals behave like humans as it is the case in fables.

Towards faithfully visualizing global linguistic diversity

Garland McNew, Curdin Derungs and Steven Moran

The most popular strategy for visualizing worldwide linguistic diversity is to utilize point symbology by plotting linguistic features as colored dots or shapes on a Mercator map projection. This approach creates illusions due to the choice of cartographic projection and also from statistical biases inherent in samples of language data and their encoding in typological databases. Here we describe these challenges and offer an approach towards faithfully visualizing linguistic diversity. Instead of Mercator, we propose an Eckert IV projection to serve as a map base layer. Instead of languages-as-points, we use Voronoi/Thiessen tessellations to model linguistic areas, including polygons for languages for which there is missing data in the sample under investigation. Lastly we discuss future work in the intersection of cartography and comparative linguistics, which must be addressed to further advance visualizations of worldwide linguistic diversity.

The GermaParl Corpus of Parliamentary Protocols

Andreas Blätte and Andre Blessing

This paper introduces the GermaParl Corpus. We outline available data, the data preparation process for preparing corpora of parliamentary debates, and the tools we used to obtained hand-coded annotations that serve as training data for classifying debates. Beyond introducing a resource that is valuable for research, we share experiences and best practices for preparing corpora of plenary protocols.

Identifying Speakers and Addressees in Dialogues Extracted from Literary Fiction

Adam Ek, Mats Wirén, Robert Östling, Kristina Nilsson Björkenstam, Gintare Grigonyte and Sofia Gustafson Capková

This paper describes an approach to identifying speakers and addressees in dialogues extracted from literary fiction, along with a dataset annotated for speaker and addressee. The overall purpose of this is to provide annotation of dialogue interaction between characters in literary corpora in order to allow for enriched search facilities and construction of social networks from the corpora. To predict speakers and addressees in a dialogue, we use a sequence labeling approach applied to a given set of characters. We use features relating to the current dialogue, the preceding narrative, and the complete preceding context. The results indicate that even with a small amount of training data, it is possible to build a fairly accurate classifier for speaker and addressee identification across different authors, though the identification of addressees is the more difficult task.

Session P11 - Lexicon (1)

9th May 2018, 14:35

Chair person: **Francesca Frontini** Poster Session

Word Embedding Evaluation Datasets and Wikipedia Title Embedding for Chinese

Chi-Yen Chen and Wei-Yun Ma

Distributed word representations are widely used in many NLP tasks, and there are lots of benchmarks to evaluate word embeddings in English. However there are barely evaluation sets with large enough amount of data for Chinese word embeddings. Therefore, in this paper, we create several evaluation sets for Chinese word embedding on both word similarity task and analogical task via translating some existing popular evaluation sets from English to Chinese. To assess the quality of translated datasets, we obtain human rating from both experts and Amazon Mechanical Turk workers. While translating the datasets, we find out that around 30 percents of word pairs in the benchmarks are Wikipedia titles. This motivate us to evaluate the performance of Wikipedia title embeddings on our new benchmarks. Thus, in this paper, not only the new benchmarks are tested but some new improved approaches of Wikipedia title embeddings are proposed. We perform training of embeddings of Wikipedia titles using not only their Wikipedia context but also their Wikipedia categories, most of categories are noun phrases, and we identify the head words of the noun phrases by a parser for further emphasizing their roles on the training of title embeddings. Experimental results and the comprehensive error analysis demonstrate that the benchmarks can precisely reflect the approaches' quality, and the effectiveness of our improved approaches on Wikipedia title embeddings are also verified and analyzed in detail.

An Automatic Learning of an Algerian Dialect Lexicon by using Multilingual Word Embeddings

ABIDI Karima and Kamel Smaili

The goal of this work consists in building automatically from a social network (Youtube) an Algerian dialect lexicon. Each entry of this lexicon is composed by a word, written in Arabic script (modern standard Arabic or dialect) or Latin script (Arabizi, French or English). To each word, several transliterations are proposed, written in a script different from the one used for the word itself. To do that, we harvested and aligned an Algerian dialect corpus by using an iterative method based on multlingual word embeddings representation. The multlinguality in the corpus is due to the fact that Algerian people use several languages to post comments in social networks: Modern Standard Arabic (MSA), Algerian dialect, French and sometimes English. In addition, the users of social networks write freely without any regard to the grammar of these languages. We tested the proposed method on a test lexicon, it leads to a score of 73% in terms of F-measure.

Candidate Ranking for Maintenance of an Online Dictionary

Claire Broad, Helen Langone and David Guy Brizan

Traditionally, the process whereby a lexicographer identifies a lexical item to add to a dictionary – a database of lexical items – has been time-consuming and subjective. In the modern age of online dictionaries, all queries for lexical entries not currently in the database are indistinguishable from a larger list of misspellings, meaning that potential new or trending entries can get lost easily. In this project, we develop a system that uses machine learning techniques to assign these "misspells" a probability of being a novel or missing entry, incorporating signals from orthography, usage by trusted online sources, and dictionary query patterns.

Language adaptation experiments via cross-lingual embeddings for related languages

Serge Sharoff

Language Adaptation (similarly to Domain Adaptation) is a general approach to extend existing resources from a better resourced language (donor) to a lesser resourced one (recipient) by exploiting the lexical and grammatical similarity between them when the two languages are related. The current study improves the state of the art in cross-lingual word embeddings by considering the impact of orthographic similarity between cognates. In particular, the use of the Weighted Levenshtein Distance combined with orthogonalisation of the translation matrix and generalised correction for hubness can considerably improve the state of the art in induction of bilingual lexicons. In addition to intrinsic evaluation in the bilingual lexicon induction task, the paper reports extrinsic evaluation of the cross-lingual embeddings via their application to the Named-Entity Recognition task across Slavonic languages. The tools and the aligned word embedding spaces for the Romance and Slavonic language families have been released.

Tools for Building an Interlinked Synonym Lexicon Network

Zdenka Uresova, Eva Fucikova, Eva Hajicova and Jan Hajic

This paper presents the structure, features and design of a new interlinked verbal synonym lexical resource called CzEngClass and the editor tool being developed to assist the work. This lexicon captures cross-lingual (Czech and English) synonyms, using valency behavior of synonymous verbs in relation to semantic roles as one of the criteria for defining such interlingual synonymy. The tool, called Synonym Class Editor - SynEd, is a user-friendly tool specifically customized to build and edit individual entries in the lexicon. It helps to keep the cross-lingual synonym classes consistent and linked to internal as well as to well-known external lexical resources. The structure of SynEd also allows to keep and edit the appropriate syntactic and semantic information for each Synonym Class member. The editor makes it possible to display examples of class members' usage in translational context in a parallel corpus. SynEd is platform independent and may be used for multiple languages. SynEd, CzEngClass and services based on them will be openly available.

Very Large-Scale Lexical Resources to Enhance Chinese and Japanese Machine Translation

Jack Halpern

A major issue in machine translation (MT) applications is the recognition and translation of named entities. This is especially true for Chinese and Japanese, whose scripts present linguistic and algorithmic challenges not found in other languages. This paper discusses some of the major issues in Japanese and Chinese MT, such as the difficulties of translating proper nouns and technical terms, and the complexities of orthographic variation in Japanese. Of special interest are neural machine translation (NMT) systems, which suffer from a serious out-of-vocabulary problem. However, the current architecture of these systems makes it technically challenging for them to alleviate this problem by supporting lexicons. This paper introduces some Very Large-Scale Lexical Resources (VLSLR) consisting of millions of named entities, and argues that the quality of MT in general, and NMT systems in particular, can be significantly enhanced through the integration of lexicons.

Combining Concepts and Their Translations from Structured Dictionaries of Uralic Minority Languages

Mika Hämäläinen, Liisa Lotta Tarvainen and Jack Rueter

Building language resources for endangered languages, especially in the case of dictionaries, requires a substantial amount of manual work. This, however, is a time-consuming undertaking, and it is also why we propose an automated method for expanding the knowledge in the existing dictionaries. In this paper, we present an approach to automatically combine conceptually divided translations from multilingual dictionaries for small Uralic languages. This is done for the noun dictionaries of Skolt Sami, Erzya, Moksha and Komi- Zyrian in such a way that the combined translations are included in the dictionaries of each language and then evaluated by professional linguists fluent in these languages. Inclusion of the method as a part of the new crowdsourced MediaWiki based pipeline for editing the dictionaries is discussed. The method can be used there not only to expand the existing dictionaries but also to provide the editors with translations when they are adding a new lexical entry to the system.

Transfer of Frames from English FrameNet to Construct Chinese FrameNet: A Bilingual Corpus-Based Approach

Tsung-Han Yang, Hen-Hsen Huang, An-Zi Yen and Hsin-Hsi Chen

Current publicly available Chinese FrameNet has a relatively low coverage of frames and lexical units compared with FrameNet in other languages. Frames are incompletely specified for some lexical units, and some critical lexical elements are even missing. That results in suitable frames cannot be triggered and filled in practical applications. This paper presents an automatic approach to constructing Chinese FrameNet. We first capture the mapping between English lexical entries and their Chinese counterparts in a large scale sentence-aligned English-Chinese bilingual corpus. Then, a semantic transfer approach is proposed based on word alignments applied to a large balanced bilingual corpus. The resource currently covers 779 frames and 36k lexical units.

EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language

Luise Dürlich and Thomas Francois

This paper introduces EFLLex, an innovative lexical resource that describes the use of 15,280 English words in pedagogical materials across the proficiency levels of the European Framework of Reference for Languages. The methodology adopted to produce the resource implies the selection of an efficient part-of-speech tagger, the use of a robust estimator for frequency computation and some manual post-editing work. The content of the resource is described and compared to other vocabulary lists (MRC and BNC) and to a reference pedagogical resource: the English Vocabulary Profile.

Session P12 - Machine Translation, SpeechToSpeech Translation (1)

9th May 2018, 14:35

Chair person: **Laurent Besacier** Poster Session

English-Basque Statistical and Neural Machine Translation

Inigo Jauregi Unanue, Lierni Garmendia Arratibel, Ehsan Zare Borzeshi and Massimo Piccardi

Neural Machine Translation (NMT) has attracted increasing attention in the recent years. However, it tends to require very large training corpora which could prove problematic for languages with low resources. For this reason, Statistical Machine Translation (SMT) continues to be a popular approach for low-resource language pairs. In this work, we address English-Basque translation and compare the performance of three contemporary statistical and neural machine translation systems: OpenNMT, Moses SMT and Google Translate. For evaluation, we employ an open-domain and an IT-domain corpora from the WMT16 resources for machine translation. In addition, we release a small dataset (Berriak) of 500 highly-accurate English-Basque translations of complex sentences useful for a thorough testing of the translation systems.

TQ-AutoTest – An Automated Test Suite for (Machine) Translation Quality

Vivien Macketanz, Renlong Ai, Aljoscha Burchardt and Hans Uszkoreit

In several areas of NLP evaluation, test suites have been used to analyze the strengths and weaknesses of systems. Today, Machine Translation (MT) quality is usually assessed by shallow automatic comparisons of MT outputs with reference corpora resulting in a number. Especially the trend towards neural MT has renewed peoples' interest in better and more analytical diagnostic methods for MT quality. In this paper we present TQ-AutoTest, a novel framework that supports a linguistic evaluation of (machine) translations using test suites. Our current test suites comprise about 5000 handcrafted test items for the language pair German–English. The framework supports the creation of tests and the semi-automatic evaluation of the MT results using regular

expressions. The expressions help to classify the results as correct, incorrect or as requiring a manual check. The approach can easily be extended to other NLP tasks where test suites can be used such as evaluating (one-shot) dialogue systems.

Exploiting Pre-Ordering for Neural Machine Translation

Yang Zhao, Jiajun Zhang and Chengqing Zong

Neural Machine Translation (NMT) has drawn much attention due to its promising translation performance in recent years. However, the under-translation and over-translation problem still remain a big challenge. Through error analysis, we find that under-translation is much more prevalent than over-translation and the source words that need to be reordered during translation are more likely to be ignored. To address the under-translation problem, we explore the pre-ordering approach for NMT. Specifically, we pre-order the source sentences to approximate the target language word order. We then combine the pre-ordering model with position embedding to enhance the monotone translation. Finally, we augment our model with the coverage mechanism to tackle the over-translation problem. Experimental results on Chinese-to-English translation have shown that our method can significantly improve the translation quality by up to 2.43 BLEU points. Furthermore, the detailed analysis demonstrates that our approach can substantially reduce the number of under-translation cases by 30.4% (compared to 17.4% using the coverage model).

Improving a Multi-Source Neural Machine Translation Model with Corpus Extension for Low-Resource Languages

Gyu Hyeon Choi, Jong Hun Shin and Young Kil Kim

In machine translation, we often try to collect resources to improve performance. However, most of the language pairs, such as Korean-Arabic and Korean-Vietnamese, do not have enough resources to train machine translation systems. In this paper, we propose the use of synthetic methods for extending a low-resource corpus and apply it to a multi-source neural machine translation model. We showed the improvement of machine translation performance through corpus extension using the synthetic method. We specifically focused on how to create source sentences that can make better target sentences, including the use of synthetic methods. We found that the corpus extension could also improve the performance of multi-source neural machine translation. We showed the corpus extension and multi-source model to be efficient methods for a low-resource language pair. Furthermore, when both methods were used together, we found better machine translation performance.

Dynamic Oracle for Neural Machine Translation in Decoding Phase

Zi-Yi Dou, Hao Zhou, Shu-Jian Huang, Xin-Yu Dai and Jia-Jun Chen

The past several years have witnessed the rapid progress of end-to-end Neural Machine Translation (NMT). However, there exists discrepancy between training and inference in NMT when decoding, which may lead to serious problems since the model might be in a part of the state space it has never seen during training. To address the issue, Scheduled Sampling has been proposed. However, there are certain limitations in Scheduled Sampling and we propose two dynamic oracle-based methods to improve it. We manage to mitigate the discrepancy by changing the training process towards a less guided scheme and meanwhile aggregating the oracle's demonstrations. Experimental results show that the proposed approaches improve translation quality over standard NMT system.

One Sentence One Model for Neural Machine Translation

Xiaoqing Li, Jiajun Zhang and Chengqing Zong

Neural machine translation (NMT) becomes a new state of the art and achieves promising translation performance using a simple encoder-decoder neural network. This neural network is trained once on the parallel corpus and the fixed network is used to translate all the test sentences. We argue that the general fixed network parameters cannot best fit each specific testing sentences. In this paper, we propose the dynamic NMT which learns a general network as usual, and then fine-tunes the network for each test sentence. The fine-tune work is done on a small set of the bilingual training data that is obtained through similarity search according to the test sentence. Extensive experiments demonstrate that this method can significantly improve the translation performance, especially when highly similar sentences are available.

A Parallel Corpus of Arabic-Japanese News Articles

Go Inoue, Nizar Habash, Yuji Matsumoto and Hiroyuki Aoyama

Much work has been done on machine translation between major language pairs including Arabic-English and English-Japanese thanks to the availability of large-scale parallel corpora with manually verified subsets of parallel sentences. However, there has been little research conducted on the Arabic-Japanese language pair due to its parallel-data scarcity, despite being a good example of interestingly contrasting differences in typology. In this paper, we describe the creation process and statistics of the Arabic-Japanese portion of the TUFS Media Corpus, a parallel corpus of translated news articles collected at Tokyo University of

Foreign Studies (TUFS). Part of the corpus is manually aligned at the sentence level for development and testing. The corpus is provided in two formats: A document-level parallel corpus in XML format, and a sentence-level parallel corpus in plain text format. We also report the first results of Arabic-Japanese phrase-based machine translation trained on our corpus.

Examining the Tip of the Iceberg: A Data Set for Idiom Translation

Marzieh Fadaee, Arianna Bisazza and Christof Monz

Neural Machine Translation (NMT) has been widely used in recent years with significant improvements for many language pairs. Although state-of-the-art NMT systems are generating progressively better translations, idiom translation remains one of the open challenges in this field. Idioms, a category of multiword expressions, are an interesting language phenomenon where the overall meaning of the expression cannot be composed from the meanings of its parts. A first important challenge is the lack of dedicated data sets for learning and evaluating idiom translation. In this paper we address this problem by creating the first large-scale data set for idiom translation. Our data set is automatically extracted from a widely used German$English translation corpus and includes, for each language direction, a targeted evaluation set where all sentences contain idioms and a regular training corpus where sentences including idioms are marked. We release this data set and use it to perform preliminary NMT experiments as the first step towards better idiom translation.

Automatic Enrichment of Terminological Resources: the IATE RDF Example

Mihael Arcan, Elena Montiel-Ponsoda, John Philip McCrae and Paul Buitelaar

Terminological resources have proven necessary in many organizations and institutions to ensure communication between experts. However, the maintenance of these resources is a very time-consuming and expensive process. Therefore, the work described in this contribution aims to automate the maintenance process of such resources. As an example, we demonstrate enriching the RDF version of IATE with new terms in the languages for which no translation was available, as well as with domain-disambiguated sentences and information about usage frequency. This is achieved by relying on machine translation trained on parallel corpora that contains the terms in question and multilingual word sense disambiguation performed on the context provided by the sentences. Our results show that for most languages translating the terms within a disambiguated context significantly outperforms the approach with randomly selected sentences.

A Comparative Study of Extremely Low-Resource Transliteration of the World's Languages

Winston Wu and David Yarowsky

Transliteration from low-resource languages is difficult, in large part due to the small amounts of data available for training transliteration systems. In this paper, we evaluate the effectiveness of several translation methods in the task of transliterating around 1000 Bible names from 591 languages into English. In this extremely low-resource task, we found that a phrase-based MT system performs much better than other methods, including a g2p system and a neural MT system. However, by combining the data and training a single neural system, we discovered significant gains over single-language systems. We release the output from each system for comparative analysis.

Translating Web Search Queries into Natural Language Questions

Adarsh Kumar, Sandipan Dandapat and Sushil Chordia

Users often query a search engine with a specific question in mind and often these queries are keywords or sub-sentential fragments. In this paper, we are proposing a method to generate well-formed natural language question from a given keyword-based query, which has the same question intent as the query.Conversion of keyword based web query into a well formed question has lots of applications in search engines, Community Question Answering (CQA) website and bots communication. We found a synergy between query-to-question problem with standard machine translation (MT) task. We have used both Statistical MT (SMT) and Neural MT(NMT) models to generate the questions from query. We have observed that MT models performs well in terms of both automatic and human evaluation.

Session P13 - Semantics (1)

9th May 2018, 14:35

Chair person: **Kyoko Kanzaki** Poster Session

Construction of a Japanese Word Similarity Dataset

Yuya Sakaizawa and Mamoru Komachi

An evaluation of distributed word representation is generally conducted using a word similarity task and/or a word analogy task. There are many datasets readily available for these tasks in English. However, evaluating distributed representation in languages that do not have such resources (e.g., Japanese) is difficult. Therefore, as a first step toward evaluating distributed representations in Japanese, we constructed a Japanese word similarity dataset. To the best of our knowledge, our dataset

is the first resource that can be used to evaluate distributed representations in Japanese. Moreover, our dataset contains various parts of speech and includes rare words in addition to common words.

Acquiring Verb Classes Through Bottom-Up Semantic Verb Clustering

Olga Majewska, Diana McCarthy, Ivan Vulić and Anna Korhonen

In this paper, we present the first analysis of bottom-up manual semantic clustering of verbs in three languages, English, Polish and Croatian. Verb classes including syntactic and semantic information have been shown to support many NLP tasks by allowing abstraction from individual words and thereby alleviating data sparseness. The availability of such classifications is however still non-existent or limited in most languages. While a range of automatic verb classification approaches have been proposed, high-quality resources and gold standards are needed for evaluation and to improve the performance of NLP systems. We investigate whether semantic verb classes in three different languages can be reliably obtained from native speakers without linguistics training. The analysis of inter-annotator agreement shows an encouraging degree of overlap in the classifications produced for each language individually, as well as across all three languages. Comparative examination of the resultant classifications provides interesting insights into cross-linguistic semantic commonalities and patterns of ambiguity.

Constructing High Quality Sense-specific Corpus and Word Embedding via Unsupervised Elimination of Pseudo Multi-sense

Haoyue Shi, Xihao Wang, Yuqi Sun and Junfeng Hu

Multi-sense word embedding is an important extension of neural word embeddings. By leveraging context of each word instance, multi-prototype version of word embeddings were accomplished to represent the multi-senses. Unfortunately, this kind of context based approach inevitably produces multiple senses which should actually be a single one, suffering from the various context of a word. Shi et al.(2016) used WordNet to evaluate the neighborhood similarity of each sense pair to detect such pseudo multi-senses. In this paper, a novel framework for unsupervised corpus sense tagging is presented, which mainly contains four steps: (a) train multi-sense word embeddings on the given corpus, using existing multi-sense word embedding frameworks; (b) detect pseudo multi-senses in the obtained embeddings, without requirement to any extra language resources; (c) label each word in the corpus with a specific sense tag, with respect to the result of pseudo multi-sense detection; (d) re-train multi-sense word embeddings with the pre-selected sense tags. We evaluate our framework by training word embeddings with the obtained sense specific corpus. On the tasks of word similarity, word analogy as well as sentence understanding, the embeddings trained on sense-specific corpus obtain better results than the basic strategy which is applied in step (a).

Urdu Word Embeddings

Samar Haider

Representing words as vectors which encode their semantic properties is an important component in natural language processing. Recent advances in distributional semantics have led to the rise of neural network-based models that use unsupervised learning to represent words as dense, distributed vectors, called 'word embeddings'. These embeddings have led to breakthroughs in performance in multiple natural language processing applications, and also hold the key to improving natural language processing for low-resource languages by helping machine learning algorithms learn patterns more easily from these richer representations of words, thereby allowing better generalization from less data. In this paper, we train the skip-gram model on more than 140 million Urdu words to create the first large-scale word embeddings for the Urdu language. We analyze the quality of the learned embeddings by looking at the closest neighbours to different words in the vector space and find that they capture a high degree of syntactic and semantic similarity between words. We evaluate this quantitatively by experimenting with different vector dimensionalities and context window sizes and measuring their performance on Urdu translations of standard word similarity tasks. The embeddings are made freely available in order to advance research on Urdu language processing.

Social Image Tags as a Source of Word Embeddings: A Task-oriented Evaluation

Mika Hasegawa, Tetsunori Kobayashi and Yoshihiko Hayashi

Distributional hypothesis has been playing a central role in statistical NLP. Recently, however, its limitation in incorporating perceptual and empirical knowledge is noted, eliciting a field of perceptually grounded computational semantics. Typical sources of features in such a research are image datasets, where images are accompanied by linguistic tags and/or descriptions. Mainstream approaches employ machine learning techniques to integrate/combine visual features with linguistic features. In contrast to or supplementing these approaches, this study assesses the effectiveness of social image tags in generating word embeddings, and argues that these generated representations exhibit somewhat different and favorable behaviors from corpus-originated representations. More specifically, we generated word embeddings by using image tags obtained from a large social

image dataset YFCC100M, which collects Flickr images and the associated tags. We evaluated the efficacy of generated word embeddings with standard semantic similarity/relatedness tasks, which showed that comparable performances with corpus-originated word embeddings were attained. These results further suggest that the generated embeddings could be effective in discriminating synonyms and antonyms, which has been an issue in distributional hypothesis-based approaches. In summary, social image tags can be utilized as yet another source of visually enforced features, provided the amount of available tags is large enough.

Towards AMR-BR: A SemBank for Brazilian Portuguese Language

Rafael Anchiêta and Thiago Pardo

We present in this paper an effort to build an AMR (Abstract Meaning Representation) annotated corpus (a semantic bank) for Brazilian Portuguese. AMR is a recent and prominent meaning representation with good acceptance and several applications in the Natural Language Processing area. Following what has been done for other languages, and using an alignment-based approach for annotation, we annotated the Little Prince book, which went into the public domain and explored some language-specific annotation issues.

Towards a Welsh Semantic Annotation System

Scott Piao, Paul Rayson, Dawn Knight and Gareth Watkins

Automatic semantic annotation of natural language data is an important task in Natural Language Processing, and a variety of semantic taggers have been developed for this task, particularly for English. However, for many languages, particularly for low-resource languages, such tools are yet to be developed. In this paper, we report on the development of an automatic Welsh semantic annotation tool (named CySemTagger) in the CorCenCC Project, which will facilitate semantic-level analysis of Welsh language data on a large scale. Based on Lancaster's USAS semantic tagger framework, this tool tags words in Welsh texts with semantic tags from a semantic classification scheme, and is designed to be compatible with multiple Welsh POS taggers and POS tagsets by mapping different tagsets into a core shared POS tagset that is used internally by CySemTagger. Our initial evaluation shows that the tagger can cover up to 91.78% of words in Welsh text. This tagger is under continuous development, and will provide a critical tool for Welsh language corpus and information processing at semantic level.

Semantic Frame Parsing for Information Extraction : the CALOR corpus

Gabriel Marzinotto, Jeremy Auguste, Frederic Bechet, Géraldine Damnati and Alexis Nasr

This paper presents a publicly available corpus of French encyclopedic history texts annotated according to the Berkeley FrameNet formalism. The main difference in our approach compared to previous works on semantic parsing with FrameNet is that we are not interested here in full text parsing but rather on partial parsing. The goal is to select from the FrameNet resources the minimal set of frames that are going to be useful for the applicative framework targeted, in our case Information Extraction from encyclopedic documents. Such an approach leverage the manual annotation of larger corpus than those obtained through full text parsing and therefore open the door to alternative methods for Frame parsing than those used so far on the FrameNet 1.5 benchmark corpus. The approaches compared in this study rely on an integrated sequence labeling model which jointly optimizes frame identification and semantic role segmentation and identification. The models compared are CRFs and multitasks bi-LSTMs.

Using a Corpus of English and Chinese Political Speeches for Metaphor Analysis

Kathleen Ahrens, Huiheng Zeng and Shun-han Rebekah Wong

In this article, we present details of our corpus of political speeches and introduce using the corpus for metaphor analysis in political discourse. Although specialized corpora on a variety of topics are now easily available, online political corpora available for public use are scarce. The database our research team has developed contains more than six million English and Chinese political speeches and is currently available free online. Researchers in many fields are able to use the multiple search functions on the website for their specific research purposes. In particular, the corpus is useful for researchers focusing on political speeches and conceptual metaphor analyses. From the perspective of metaphor study, we have taken advantage of several functions to facilitate the corpus-based metaphor analyses. In short, this database enriches the current bilingual resources and contributes to the evaluation of political language by linguists and political scientists.

A Multi- versus a Single-classifier Approach for the Identification of Modality in the Portuguese Language

João Sequeira, Teresa Gonçalves, Paulo Quaresma, Amália Mendes and Iris Hendrickx

This work presents a comparative study between two different approaches to build an automatic classification system for

Modality values in the Portuguese language. One approach uses a single multi-class classifier with the full dataset that includes eleven modal verbs; the other builds different classifiers, one for each verb. The performance is measured using precision, recall and F 1 . Due to the unbalanced nature of the dataset a weighted average approach was calculated for each metric. We use support vector machines as our classifier and experimented with various SVM kernels to find the optimal classifier for the task at hand. We experimented with several different types of feature attributes representing parse tree information and compare these complex feature representation against a simple bag-of-words feature representation as baseline. The best obtained F 1 values are above 0.60 and from the results it is possible to conclude that there is no significant difference between both approaches.

Session P14 - Word Sense Disambiguation

9th May 2018, 14:35

Chair person: **Maite Melero** Poster Session

All-words Word Sense Disambiguation Using Concept Embeddings

Rui Suzuki, Kanako Komiya, Masayuki Asahara, Minoru Sasaki and Hiroyuki Shinnou

All-words word sense disambiguation (all-words WSD) is the task of identifying the senses of all words in a document. Since the sense of a word depends on the context, such as the surrounding words, similar words are believed to have similar sets of surrounding words. We therefore predict the target word senses by calculating the distances between the surrounding word vectors of the target words and their synonyms using word embeddings. In addition, we introduce the new idea of concept embeddings, constructed from concept tag sequences created from the results of previous prediction steps. We predict the target word senses using the distances between surrounding word vectors constructed from word and concept embeddings, via a bootstrapped iterative process. Experimental results show that these concept embeddings were able to improve the performance of Japanese all-words WSD.

Enhancing Modern Supervised Word Sense Disambiguation Models by Semantic Lexical Resources

Stefano Melacci, Achille Globo and Leonardo Rigutini

Supervised models for Word Sense Disambiguation (WSD) currently yield to state-of-the-art results in the most popular benchmarks. Despite the recent introduction of Word Embeddings and Recurrent Neural Networks to design powerful context-related features, the interest in improving WSD models using Semantic Lexical Resources (SLRs) is mostly restricted to knowledge-based approaches. In this paper, we enhance "modern" supervised WSD models exploiting two popular SLRs: WordNet and WordNet Domains. We propose an effective way to introduce semantic features into the classifiers, and we consider using the SLR structure to augment the training data. We study the effect of different types of semantic features, investigating their interaction with local contexts encoded by means of mixtures of Word Embeddings or Recurrent Neural Networks, and we extend the proposed model into a novel multi-layer architecture for WSD. A detailed experimental comparison in the recent Unified Evaluation Framework (Raganato et al., 2017) shows that the proposed approach leads to supervised models that compare favourably with the state-of-the art.

An Unsupervised Word Sense Disambiguation System for Under-Resourced Languages

Dmitry Ustalov, Denis Teslenko, Alexander Panchenko, Mikhail Chersnoskutov, Chris Biemann and Simone Paolo Ponzetto

In this paper, we present Watasense, an unsupervised system for word sense disambiguation. Given a sentence, the system chooses the most relevant sense of each input word with respect to the semantic similarity between the given sentence and the synset constituting the sense of the target word. Watasense has two modes of operation. The sparse mode uses the traditional vector space model to estimate the most similar word sense corresponding to its context. The dense mode, instead, uses synset embeddings to cope with the sparsity problem. We describe the architecture of the present system and also conduct its evaluation on three different lexical semantic resources for Russian. We found that the dense mode substantially outperforms the sparse one on all datasets according to the adjusted Rand index.

Unsupervised Korean Word Sense Disambiguation using CoreNet

Kijong Han, Sangha Nam, Jiseong Kim, Younggyun Hahm and KEY-SUN CHOI

In this study, we investigated unsupervised learning based Korean word sense disambiguation (WSD) using CoreNet, a Korean lexical semantic network. To facilitate the application of WSD to practical natural language processing problems, a reasonable method is required to distinguish between sense candidates. We therefore performed coarse-grained Korean WSD studies while utilizing the hierarchical semantic categories of CoreNet to distinguish between sense candidates. In our unsupervised approach, we applied a knowledge-based model that incorporated a Markov random field and dependency parsing to the Korean language in addition to utilizing the semantic categories of CoreNet. Our experimental results demonstrate that the developed

CoreNet based coarse-grained WSD technique exhibited an 80.9% accuracy on the datasets we constructed, and was proven to be effective for practical applications.

UFSAC: Unification of Sense Annotated Corpora and Tools

Loïc Vial, Benjamin Lecouteux and Didier Schwab

In Word Sense Disambiguation, sense annotated corpora are often essential for evaluating a system and also valuable in order to reach a good efficiency. Always created for a specific purpose, there are today a dozen of sense annotated English corpora, in various formats and using different versions of WordNet. The main hypothesis of this work is that it should be possible to build a disambiguation system by using any of these corpora during the training phase or during the testing phase regardless of their original purpose. In this article, we present UFSAC: a format of corpus that can be used for either training or testing a disambiguation system, and the process we followed for constructing this format. We give to the community the whole set of sense annotated English corpora that we know, in this unified format, when the copyright allows it, with sense keys converted to the last version of WordNet. We also provide the source code for building these corpora from their original data, and a complete Java API for manipulating corpora in this format. The whole resource is available at the following URL: https://github.com/getalp/UFSAC.

Retrofitting Word Representations for Unsupervised Sense Aware Word Similarities

Steffen Remus and Chris Biemann

Standard word embeddings lack the possibility to distinguish senses of a word by projecting them to exactly one vector. This has a negative effect particularly when computing similarity scores between words using standard vector-based similarity measures such as cosine similarity. We argue that minor senses play an important role in word similarity computations, hence we use an unsupervised sense inventory resource to retrofit monolingual word embeddings, producing sense-aware embeddings. Using retrofitted sense-aware embeddings, we show improved word similarity and relatedness results on multiple word embeddings and multiple established word similarity tasks, sometimes up to an impressive margin of 0.15 Spearman correlation score.

FastSense: An Efficient Word Sense Disambiguation Classifier

Tolga Uslu, Alexander Mehler, Daniel Baumartz, Alexander Henlein and Wahed Hemati

The task of Word Sense Disambiguation (WSD) is to determine the meaning of an ambiguous word in a given context. In spite of its importance for most NLP pipelines, WSD can still be seen to be unsolved. The reason is that we currently lack tools for WSD that handle big data – "big" in terms of the number of ambiguous words and in terms of the overall number of senses to be distinguished. This desideratum is exactly the objective of fastSense, an efficient neural network-based tool for word sense disambiguation introduced in this paper. We train and test fastSense by means of the disambiguation pages of the German Wikipedia. In addition, we evaluate fastSense in the context of Senseval and SemEval. By reference to Senseval and SemEval we additionally perform a parameter study. We show that fastSense can process huge amounts of data quickly and also surpasses state-of-the-art tools in terms of F-measure.

Session O9 - Bio-medical Corpora

9th May 2018, 16:35

Chair person: **Paul Rayson** Oral Session

A FrameNet for Cancer Information in Clinical Narratives: Schema and Annotation

Kirk Roberts, Yuqi Si, Anshul Gandhi and Elmer Bernstam

This paper presents a pilot project named Cancer FrameNet. The project's goal is a general-purpose natural language processing (NLP) resource for cancer-related information in clinical notes (i.e., patient records in an electronic health record system). While previous cancer NLP annotation projects have largely been ad hoc resources to address a specific and immediate information need, the frame semantic method employed here emphasizes the information presented in the notes themselves and its linguistic structure. To this end, three semantic frames (targeting the high-level tasks of cancer diagnoses, cancer therapeutic procedures, and tumor descriptions) are created and annotated on a clinical text corpus. Prior to annotation, candidate sentences are extracted from a clinical data warehouse and de-identified to remove any private information. The frames are then annotated with the three frames totaling over thirty frame elements. This paper describes these steps in the pilot project and discusses issues encountered to evaluate the feasibility of general-purpose linguistic resources for extracting cancer-related information.

A New Corpus to Support Text Mining for the Curation of Metabolites in the ChEBI Database

Matthew Shardlow, Nhung Nguyen, Gareth Owen, Claire O'Donovan, Andrew Leach, John McNaught, Steve Turner and Sophia Ananiadou

We present a new corpus of 200 abstracts and 100 full text papers which have been annotated with named entities and relations in the biomedical domain as part of the OpenMinTeD project. This

corpus facilitates the goal in OpenMinTeD of making text and data mining accessible to the users who need it most. We describe the process we took to annotate the corpus with entities (Metabolite, Chemical, Protein, Species, Biological Activity and Spectral Data) and relations (Isolated From, Associated With, Binds With and Metabolite Of). We report inter-annotator agreement (using F-score) for entities of between 0.796 and 0.892 using a strict matching protocol and between 0.875 and 0.963 using a relaxed matching protocol. For relations we report inter annotator agreement of between 0.591 and 0.693 using a strict matching protocol and between 0.744 and 0.793 using a relaxed matching protocol. We describe how this corpus can be used within ChEBI to facilitate text and data mining and how the integration of this work with the OpenMinTeD text and data mining platform will aid curation of ChEBI and other biomedical databases.

Parallel Corpora for the Biomedical Domain

Aurélie Névéol, Antonio Jimeno Yepes, Mariana Neves and Karin Verspoor

A vast amount of biomedical information is available in the form of scientific literature and government-authored patient information documents. While English is the most widely used language in many of these sources, there is a need to provide access to health information in languages other than English. Parallel corpora can be leveraged to implement cross-lingual information retrieval or machine translation tools. Herein, we review the extent of parallel corpus coverage in the biomedical domain. Specifically, we perform a scoping review of existing resources and we describe the recent development of new datasets for scientific literature (the EDP dataset and an extension of the Scielo corpus) and clinical trials (the ReBEC corpus). These corpora are currently being used in the biomedical task in the Conference on Machine Translation (WMT'16 and WMT'17), which illustrates their potential for improving and evaluating biomedical machine translation systems. Furthermore, we suggest additional applications for multilingual natural language processing using these resources, and plan to extend resource coverage to additional text genres and language pairs.

Medical Entity Corpus with PICO elements and Sentiment Analysis

Markus Zlabinger, Linda Andersson, Allan Hanbury, Michael Andersson, Vanessa Quasnik and Jon Brassey

In this paper, we present our process to establish a PICO and a sentiment annotated corpus of clinical trial publications. PICO stands for Population, Intervention, Comparison and Outcome — these four classes can be used for more advanced and specific search queries. For example, a physician can determine how well a drug works only in the subgroup of children. Additionally to the PICO extraction, we conducted a sentiment annotation, where the sentiment refers to whether the conclusion of a trial was positive, negative or neutral. We created both corpora with the help of medical experts and non-experts as annotators.

Session O10 - MultiWord Expressions

9th May 2018, 16:35

Chair person: **Francis Bond** Oral Session

Word Embedding Approach for Synonym Extraction of Multi-Word Terms

Amir Hazem and Béatrice Daille

The acquisition of synonyms and quasi-synonyms of multi-word terms (MWTs) is a relatively new and under represented topic of research. However, dealing with MWT synonyms and semantically related terms is a challenging task, especially when MWT synonyms are single word terms (SWTs) or MWTs of different lengths. While several researches addressed synonym extraction of SWTs, few of them dealt with MWTs and fewer or none while MWTs synonyms are of variable lengths. The present research aims at introducing a new word-embedding-based approach for the automatic acquisition of synonyms of MWTs that manage length variability. We evaluate our approach on two specialized domain corpora, a French/English corpus of the wind energy domain and a French/English corpus of the breast cancer domain and show superior results compared to baseline approaches.

A Large Automatically-Acquired All-Words List of Multiword Expressions Scored for Compositionality

Will Roberts and Markus Egg

We present and make available a large automatically-acquired all-words list of English multiword expressions scored for compositionality. Intrinsic evaluation against manually-produced gold standards demonstrates that our compositionality estimates are sound, and extrinsic evaluation via incorporation of our list into a machine translation system to better handle idiomatic expressions results in a statistically significant improvement to the system's BLEU scores. As the method used to produce the list is language-independent, we also make available lists in seven other European languages.

A Hybrid Approach for Automatic Extraction of Bilingual Multiword Expressions from Parallel Corpora

Nasredine Semmar

Specific-domain bilingual lexicons play an important role for domain adaptation in machine translation. The entries of these types of lexicons are mostly composed of MultiWord Expressions (MWEs). The manual construction of MWEs bilingual lexicons is costly and time-consuming. We often use word alignment approaches to automatically construct bilingual lexicons of MWEs from parallel corpora. We present in this paper a hybrid approach to extract and align MWEs from parallel corpora in a one-step process. We formalize the alignment process as an integer linear programming problem in order to find an approximated optimal solution. This process generates lists of MWEs with their translations, which are then filtered using linguistic patterns for the construction of the bilingual lexicons of MWEs. We evaluate the bilingual lexicons of MWEs produced by this approach using two methods: a manual evaluation of the alignment quality and an evaluation of the impact of this alignment on the translation quality of the phrase-based statistical machine translation system Moses. We experimentally show that the integration of the bilingual MWEs and their linguistic information into the translation model improves the performance of Moses.

No more beating about the bush : A Step towards Idiom Handling for Indian Language NLP

Ruchit Agrawal, Vighnesh Chenthil Kumar, Vigneshwaran Muralidaran and Dipti Sharma

One of the major challenges in the field of Natural Language Processing (NLP) is the handling of idioms; seemingly ordinary phrases which could be further conjugated or even spread across the sentence to fit the context. Since idioms are a part of natural language, the ability to tackle them brings us closer to creating efficient NLP tools. This paper presents a multilingual parallel idiom dataset for seven Indian languages in addition to English and demonstrates its usefulness for two NLP applications - Machine Translation and Sentiment Analysis. We observe significant improvement for both the subtasks over baseline models trained without employing the idiom dataset.

Session O11 - Time & Space

9th May 2018, 16:35

Chair person: **Kyoko Ohara** Oral Session

Sentence Level Temporality Detection using an Implicit Time-sensed Resource

Sabyasachi Kamila, Asif Ekbal and Pushpak Bhattacharyya

Temporal sense detection of any word is an important aspect for detecting temporality at the sentence level. In this paper, at first, we build a temporal resource based on a semi-supervised learning approach where each Hindi-WordNet synset is classified into one of the five classes, namely past, present, future, neutral and atemporal. This resource is then utilized for tagging the sentences with past, present and future temporal senses. For the sentence-level tagging, we use a rule-based as well as a machine learning-based approach. We provide detailed analysis along with necessary resources.

Comprehensive Annotation of Various Types of Temporal Information on the Time Axis

Tomohiro Sakaguchi, Daisuke Kawahara and Sadao Kurohashi

In order to make the temporal interpretation of text, there have been many studies linking event and temporal information, such as temporal ordering of events and timeline generation. To train and evaluate models in these studies, many corpora that associate event information with time information have been developed. In this paper, we propose an annotation scheme that anchors expressions in text to the time axis comprehensively, extending the previous studies in the following two points. One of the points is to annotate not only expressions with strong temporality but also expressions with weak temporality, such as states and habits. The other point is that various types of temporal information, such as frequency and duration, can be anchored to the time axis. Using this annotation scheme, we annotated a subset of Kyoto University Text Corpus. Since the corpus has already been annotated predicate-argument structures and coreference relations, it can be utilized for integrated information analysis of events, entities and time.

Systems' Agreements and Disagreements in Temporal Processing: An Extensive Error Analysis of the TempEval-3 Task

Tommaso Caselli and Roser Morante

In this article we review Temporal Processing systems that participated in the TempEval-3 task as a basis to develop our own

system, that we also present and release. The system incorporates high level lexical semantic features, obtaining the best scores for event detection (F1-Class 72.24) and second best result for temporal relation classification from raw text (F1 29.69) when evaluated on the TempEval-3 data. Additionally, we analyse the errors of all TempEval-3 systems for which the output is publicly available with the purpose of finding out what are the weaknesses of current approaches. Although incorporating lexical semantics features increases the performance of our system, the error analysis shows that systems should incorporate inference mechanisms and world knowledge, as well as having strategies to compensate for data skewness.

Annotating Temporally-Anchored Spatial Knowledge by Leveraging Syntactic Dependencies

Alakananda Vempala and Eduardo Blanco

This paper presents a two-step methodology to annotate temporally-anchored spatial knowledge on top of OntoNotes. We first generate potential knowledge using syntactic dependencies, and then crowdsource annotations to validate the potential knowledge. The resulting annotations indicate how long entities are or are not located somewhere, and temporally anchor this information. Crowdsourcing experiments show that spatial inferences are ubiquitous and intuitive, and experimental results show that they can be done automatically.

Session O12 - Computer Assisted Language Learning

9th May 2018, 16:35

Chair person: **Zygmunt Vetulani** Oral Session

Contextualized Usage-Based Material Selection

Dirk De Hertog and Piet Desmet

In this paper, we combine several NLP-functionalities to organize examples drawn from corpora. The application's primary target audience are language learners. Currently, authentic linguistic examples for a given keyword search are often organized alphabetically according to context. From this, it is not always clear which contextual regularities actually exist on a syntactic, collocational and semantic level. Showing information at different levels of abstraction will help with the discovery of linguistic regularities and thus improve linguistic understanding. Practically this translates in a system that groups retrieved results on syntactic grounds, after which the examples are further organized at the hand of semantic similarity within certain phrasal slots. Visualization algorithms are then used to show focused information in phrasal slots, laying bare semantic restrictions within the construction.

CBFC: a parallel L2 speech corpus for Korean and French learners

Hiyon Yoo and Inyoung Kim

In this paper, we present the design of a bilingual corpus of French learners of Korean and Korean learners of French using the same experimental design. This language resource contains mainly speech data, gathered among learners with different proficiency levels and in different speaking contexts (read and spontaneous speech). We aim at providing a translated and annotated corpus to the scientific community which can be used for a large array of purposes in the field of theoretical but also applied linguistics.

SW4ALL: a CEFR Classified and Aligned Corpus for Language Learning

Rodrigo Wilkens, Leonardo Zilio and Cédrick Fairon

Learning a second language is a task that requires a good amount of time and dedication. Part of the process involves the reading and writing of texts in the target language, and so, to facilitate this process, especially in terms of reading, teachers tend to search for texts that are associated to the interests and capabilities of the learners. But the search for this kind of text is also a time-consuming task. By focusing on this need for texts that are suited for different language learners, we present in this study the SW4ALL, a corpus with documents classified by language proficiency level (based on the CEFR recommendations) that allows the learner to observe ways of describing the same topic or content by using strategies from different proficiency levels. This corpus uses the alignments between the English Wikipedia and the Simple English Wikipedia for ensuring the use of similar content or topic in pairs of text, and an annotation of language levels for ensuring the difference of language proficiency level between them. Considering the size of the corpus, we used an automatic approach for the annotation, followed by an analysis to sort out annotation errors. SW4ALL contains 8.669 pairs of documents that present different levels of language proficiency.

Towards a Diagnosis of Textual Difficulties for Children with Dyslexia

Solen Quiniou and Béatrice Daille

Children's books are generally designed for children of a certain age group. For underage children or children with reading disorders, like dyslexia, there may be passages of the books that are difficult to understand. This can be due to words not known in the vocabulary of underage children, to words made of complex subparts (to pronounce, for example), or to the presence of anaphoras that have to be resolved by the children during the reading. In this paper, we present a study on diagnosing the difficulties appearing in French children's books.

We are more particularly interested on the difficulties coming from pronouns that can disrupt the story comprehension for children with dyslexia and we focus on the subject pronouns "il" and "elle" (corresponding to the pronoun "it"). We automatically identify the pleonastic pronouns (eg, in "it's raining") and the pronominal anaphoras, as well as the referents of the pronominal anaphoras. We also detect difficult anaphoras that are more likely to lead to miscomprehension from the children: this is the first step to diagnose the textual difficulties of children's books. We evaluate our approach on several French children's books that were manually annotated by a speech therapist. Our first results show that we are able to detect half of the difficult anaphorical pronouns.

Session P15 - Annotation Methods and Tools

9th May 2018, 16:35

Chair person: **Ron Artstein** Poster Session

Text Annotation Graphs: Annotating Complex Natural Language Phenomena

Angus Forbes, Kristine Lee, Gus Hahn-Powell, Marco A. Valenzuela-Escarcega and Mihai Surdeanu

This paper introduces a new web-based software tool for annotating text, Text Annotation Graphs, or TAG. It provides functionality for representing complex relationships between words and word phrases that are not available in other software tools, including the ability to define and visualize relationships between the relationships themselves (semantic hypergraphs). Additionally, we include an approach to representing text annotations in which annotation subgraphs, or semantic summaries, are used to show relationships outside of the sequential context of the text itself. Users can use these subgraphs to quickly find similar structures within the current document or external annotated documents. Initially, TAG was developed to support information extraction tasks on a large database of biomedical articles. However, our software is flexible enough to support a wide range of annotation tasks for many domains. Examples are provided that showcase TAG's capabilities on morphological parsing and event extraction tasks.

Manzanilla: An Image Annotation Tool for TKB Building

Arianne Reimerink and Pilar León-Araúz

Much has been written regarding the importance of combining visual and textual information to enhance knowledge acquisition (Paivio, 1971, 1986; Mayer & Anderson, 1992). However, the combination of images and text still needs further analysis (Faber, 2012; Prieto, 2008; Prieto & Faber, 2012). An in-depth analysis of the features of images provides the means to develop selection criteria for specific representation purposes. The combination of conceptual content, image type based on morphological characteristics, and functional criteria can be used to enhance the selection and annotation of images that explicitly focus on the conceptual propositions that best define concepts in a knowledge base. Manzanilla is an image annotation tool specifically created for EcoLexicon, a multilingual and multimodal terminological knowledge base (TKB) on the environment. It is powered by Camomile (Poignant et al., 2016) according to the selection and annotation criteria resulting from ten years of research on multimodality within the framework of Frame-Based Terminology (FBT; Faber, León-Araúz & Reimerink, 2014). The tool was created to enhance the consistency of knowledge representation through images with the conceptual knowledge in EcoLexicon and to improve image reusability.

Tools for The Production of Analogical Grids and a Resource of N-gram Analogical Grids in 11 Languages

Rashel Fam and Yves Lepage

We release a Python module containing several tools to build analogical grids from words contained in a corpus. The module implements several previously presented algorithms. The tools are language-independent. This permits their use with any language and any writing system. We hope that the tools will ease research in morphology by allowing researchers to automatically obtain structured representations of the vocabulary contained in corpora or linguistic data. We also release analogical grids built on the vocabularies contained in 1,000 corresponding lines of the 11 different language versions of the Europarl corpus v.3. The grids were built on N-grams of different lengths, from words to 6-grams. We hope that the use of structured parallel data will foster research in comparative linguistics.

The Automatic Annotation of the Semiotic Type of Hand Gestures in Obama' s Humorous Speeches

Costanza Navarretta

This paper describes a pilot study act to investigate the semiotic types of hand gestures in video-recorded speeches and their automatic classification. Gestures, which also comprise e.g. head movements and body posture, contribute to the successful delivery of the message by reinforcing what is expressed by speech or by adding new information to what is uttered. The automatic classification of the semiotic type of gestures from their shape description can contribute to their interpretation

in human-human communication and in advanced multimodal interactive systems. We annotated and analysed hand gestures produced by Barack Obama during two speeches at the Annual White House Correspondent Dinners and found differences in the contexts in which various hand gesture types were used. Then, we trained machine learning algorithms to classify the semiotic type of the hand gestures. The F-score obtained by the best performing algorithm on the classification of four semiotic types is 0.59. Surprisingly, the shape feature that contributes mostly to classification is the trajectory of the left hand. The results of this study are promising, but they should be tested on more data of different type, produced by different speakers and in more languages.

WASA: A Web Application for Sequence Annotation

Fahad AlGhamdi and Mona Diab

Data annotation is an important and necessary task for all NLP applications. Designing and implementing a web-based application that enables many annotators to annotate and enter their input into one central database is not a trivial task. These kinds of web-based applications require a consistent and robust backup for the underlying database and support to enhance the efficiency and speed of the annotation. Also, they need to ensure that the annotations are stored with a minimal amount of redundancy in order to take advantage of the available resources(e.g, storage space). In this paper, we introduce WASA, a web-based annotation system for managing large-scale multilingual Code Switching (CS) data annotation. Although WASA has the ability to perform the annotation for any token sequence with arbitrary tag sets, we will focus on how WASA is used for CS annotation. The system supports concurrent annotation, handles multiple encodings, allows for several levels of management control, and enables quality control measures while seamlessly reporting annotation statistics from various perspectives and at different levels of granularity. Moreover, the system is integrated with a robust language specific date prepossessing tool to enhance the speed and efficiency of the annotation. We describe the annotation and the administration interfaces as well as the backend engine.

Annotation and Quantitative Analysis of Speaker Information in Novel Conversation Sentences in Japanese

Makoto Yamazaki, Yumi Miyazaki and Wakako Kashino

This study undertook a quantitative lexicological analysis using attributed speaker information, and reports on the problems of creating standards when annotating speaker information (gender and age) of conversation sentences in novels. In this paper, we performed a comparative analysis of vocabulary use by gender and age of conversation sentences and descriptive part sentences, as well as on the differences between Japanese novels and translations of foreign novels. In addition, a comparison with other spoken language materials was made.

PDFAnno: a Web-based Linguistic Annotation Tool for PDF Documents

Hiroyuki Shindo, Yohei Munesada and Yuji Matsumoto

We present PDFAnno, a web-based linguistic annotation tool for PDF documents. PDF has become widespread standard for various types of publications, however, current tools for linguistic annotation mostly focus on plain-text documents. PDFAnno offers functions for various types of linguistic annotations directly on PDF, including named entity, dependency relation, and coreference chain. Furthermore, for multi-user support, it allows simultaneous visualization of multi-user's annotations on the single PDF, which is useful for checking inter-annotator agreement and resolving annotation conflicts. PDFAnno is freely available under open-source license at https://github.com/paperai/pdfanno.

A Lightweight Modeling Middleware for Corpus Processing

Markus Gärtner and Jonas Kuhn

Present-day empirical research in computational or theoretical linguistics has at its disposal an enormous wealth in the form of richly annotated and diverse corpus resources. Especially the points of contact between modalities are areas of exciting new research. However, progress in those areas in particular suffers from poor coverage in terms of visualization or query systems. Many limitations for such tools stem from the non-uniform representations of very diverse resources and the lack of standards that address this problem from the perspective of processing or querying. In this paper we present our framework for modeling arbitrary multi-modal corpus resources in a unified form for processing tools. It serves as a middleware system and combines the expressiveness of general graph-based models with a rich metadata schema to preserve linguistic specificity. By separating data structures and their linguistic interpretations, it assists tools on top of it so that they can in turn allow their users to more efficiently exploit corpus resources.

An Annotation Language for Semantic Search of Legal Sources

Adeline Nazarenko, Francois Levy and Adam Wyner

While formalizing legal sources is an important challenge, the generation of a formal representation from legal texts has been far less considered and requires considerable expertise. In order

to improve the uniformity, richness, and efficiency of legal annotation, it is necessary to experiment with annotations and the annotation process. This paper reports on a first experiment, which was a campaign to annotate legal instruments provided by the Scottish Government's Parliamentary Counsel Office and bearing on Scottish smoking legislation and regulation. A small set of elements related to LegalRuleML was used. An initial guideline manual was produced to annotate the text using annotations related to these elements. The resulting annotated corpus is converted into a LegalRuleML XML compliant document, then made available via an online visualisation and query tool. In the course of annotating the documents, a range of important interpretive and practical issues arose, highlighting the value of a focused study on legal text annotation.

Resource Interoperability for Sustainable Benchmarking: The Case of Events

Chantal Van Son, Oana Inel, Roser Morante, Lora Aroyo and Piek Vossen

With the continuous growth of benchmark corpora, which often annotate the same documents, there is a range of opportunities to compare and combine similar and complementary annotations. However, these opportunities are hampered by a wide range of problems that are related to the lack of resource interoperability. In this paper, we illustrate these problems by assessing aspects of interoperability at the document-level across a set of 20 corpora annotated with (aspects of) events. The issues range from applying different document naming conventions, to mismatches in textual content and structural/conceptual differences among annotation schemes. We provide insight into the exact document intersections between the corpora by mapping their document identifiers and perform an empirical analysis of event annotations showing their compatibility and consistency in and across the corpora. This way, we aim to make the community more aware of the challenges and opportunities and to inspire working collaboratively towards interoperable resources.

Parsivar: A Language Processing Toolkit for Persian

Salar Mohtaj, Behnam Roshanfekr, Atefeh Zafarian and Habibollah Asghari

With the growth of Internet usage, a massive amount of textual data is generated on social media and the Web. As the text on the Web are generated by different authors with various types of writing styles and different encodings, a preprocessing step is required before applying any NLP task. The goal of preprocessing is to convert text into a standard format that makes it easy to extract information from documents and sentences. Moreover, the problem is more acute when we deal with Arabic script-based languages, in which there are some different kinds of encoding schemes, different kinds of writing styles and the spaces between or within the words. This paper introduces a preprocessing toolkit named as Parsivar, which is a comprehensive set of tools for Persian text preprocessing tasks. This toolkit performs various kinds of activities comprised of normalization, space correction, tokenization, stemming, parts of speech tagging and shallow parsing. To evaluate the performance of the proposed toolkit, both intrinsic and extrinsic approaches for evaluation have been applied. A Persian plagiarism detection system has been exploited as a downstream task for extrinsic evaluation of the proposed toolkit. The results have revealed that our toolkit outperforms the available Persian preprocessing toolkits by about 8 percent in terms of F1.

Multilingual Word Segmentation: Training Many Language-Specific Tokenizers Smoothly Thanks to the Universal Dependencies Corpus

Erwan Moreau and Carl Vogel

This paper describes how a tokenizer can be trained from any dataset in the Universal Dependencies 2.1 corpus. A software tool, which relies on Elephant to perform the training, is also made available. Beyond providing the community with a large choice of language-specific tokenizers, we argue in this paper that: (1) tokenization should be considered as a supervised task; (2) language scalability requires a streamlined software engineering process across languages.

Build Fast and Accurate Lemmatization for Arabic

Hamdy Mubarak

In this paper we describe the complexity of building a lemmatizer for Arabic which has a rich and complex morphology, and show some differences between lemmatization and surface stemming, i.e. removing prefixes and suffixes from words. We discuss the need for a fast and accurate lammatization to enhance Arabic Information Retrieval results. We also introduce a new dataset that can be used to test lemmatization accuracy, and an efficient lemmatization algorithm that outperforms state-of-the-art Arabic lemmatization in terms of accuracy and speed. We share the dataset and the code for research purposes.

Session P16 - Corpus Creation, Annotation, Use (1)

9th May 2018, 16:35

Chair person: **Prokopis Prokopidis** Poster Session

JESC: Japanese-English Subtitle Corpus

Reid Pryzant, Youngjoo Chung, Dan Jurafsky and Denny Britz

In this paper we describe the Japanese-English Subtitle Corpus (JESC). JESC is a large Japanese-English parallel corpus covering the underrepresented domain of conversational dialogue. It consists of more than 3.2 million examples, making it the largest freely available dataset of its kind. The corpus was assembled by crawling and aligning subtitles found on the web. The assembly process incorporates a number of novel preprocessing elements to ensure high monolingual fluency and accurate bilingual alignments. We summarize its contents and evaluate its quality using human experts and baseline machine translation (MT) systems.

Building a Corpus for Personality-dependent Natural Language Understanding and Generation

Ricelli Ramos, Georges Neto, Barbara Silva, Danielle Monteiro, Ivandré Paraboni and Rafael Dias

The computational treatment of human personality - both for the recognition of personality traits from text and for the generation of text so as to reflect a particular set of traits - is central to the development of NLP applications. As a means to provide a basic resource for studies of this kind, this article describes the b5 corpus, a collection of controlled and free (non-topic specific) texts produced in different (e.g., referential or descriptive) communicative tasks, and accompanied by inventories of personality of their authors and additional demographics. The present discussion is mainly focused on the various corpus components and on the data collection task itself, but preliminary results of personality recognition from text are presented in order to illustrate how the corpus data may be reused. The b5 corpus aims to provide support for a wide range of NLP studies based on personality information and it is, to the best of our knowledge, the largest resource of this kind to be made available for research purposes in the Brazilian Portuguese language.

Linguistic and Sociolinguistic Annotation of 17th Century Dutch Letters

Marijn Schraagen, Feike Dietz and Marjo Van Koppen

Developments in the Dutch language during the 17th century, part of the Early Modern period, form an active research topic in historical linguistics and literature. To enable automatic quantitative analysis, a corpus of letters by the 17th century Dutch author and politician P.C. Hooft is manually annotated with parts-of-speech, document segmentation and sociolinguistic metadata. The corpus is developed as part of the Nederlab online research portal, which is available through the CLARIN ERIC European research infrastructure. This paper discusses the design and evaluation of the annotation effort, as well as adding new annotations to an existing annotated corpus.

Simplified Corpus with Core Vocabulary

Takumi Maruyama and Kazuhide Yamamoto

We have constructed the simplified corpus for the Japanese language and selected the core vocabulary. The corpus has 50,000 manually simplified and aligned sentences. This corpus contains the original sentences, simplified sentences and English translation of the original sentences. It can be used for automatic text simplification as well as translating simple Japanese into English and vice-versa. The core vocabulary is restricted to 2,000 words where it is selected by accounting for several factors such as meaning preservation, variation, simplicity and the UniDic word segmentation criterion. We repeated the construction of the simplified corpus and, subsequently, updated the core vocabulary accordingly. As a result, despite vocabulary restrictions, our corpus achieved high quality in grammaticality and meaning preservation. In addition to representing a wide range of expressions, the core vocabulary's limited number helped in showing similarities of expressions among simplified sentences. We believe that the same quality can be obtained by extending this corpus.

A Pragmatic Approach for Classical Chinese Word Segmentation

Shilei Huang and Jiangqin Wu

Word segmentation, a fundamental technology for lots of downstream applications, plays a significant role in Natural Language Processing, especially for those languages without explicit delimiters, like Chinese, Korean, Japanese and etc. Basically, word segmentation for modern Chinese is worked out to a certain extent. Nevertheless, Classical Chinese is largely neglected, mainly owing to its obsoleteness. One of the biggest problems for the researches of Classical Chinese word segmentation (CCWS) is lacking in standard large-scale shareable marked-up corpora, for the fact that the most excellent approaches, solving word segmentation, are based on machine learning or statistical methods which need quality-assured marked-up corpora. In this paper, we propose a pragmatic approach founded on the difference of t-score (dts) and Baidu Baike (the largest Chinese-language encyclopedia like Wikipedia) in order to deal with CCWS without any marked-up corpus. We extract candidate

words as well as their corresponding frequency from the Twenty-Five Histories (Twenty-Four Histories and Draft History of Qing) to build a lexicon, and conduct segmentation experiments with it. The F-Score of our approach on the whole evaluation data set is 76.84%. Compared with traditional collocation-based methods, ours makes the segmentation more accurate.

ASAP++: Enriching the ASAP Automated Essay Grading Dataset with Essay Attribute Scores

Sandeep Mathias and Pushpak Bhattacharyya

In this paper, we describe the creation of a resource - ASAP++ - which is basically annotations of the Automatic Student Assessment Prize's Automatic Essay Grading dataset. These annotations are scores for different attributes of the essays, such as content, word choice, organization, sentence fluency, etc. Each of these essays is scored by an annotator. We also report the results of each of the attributes using a Random Forest Classifier using a baseline set of task independent features. We release and share this resource to facilitate further research into these attributes of essay grading.

MirasText: An Automatically Generated Text Corpus for Persian

Behnam Sabeti, Hossein Abedi Firouzjaee, Ali Janalizadeh Choobbasti, Seyed hani elamahdi Mortazavi Najafabadi and Amir Vaheb

Natural Language Processing is one of the most important fields of artificial intelligence. The rapid growth of digital content has made this field both practical and challenging at the same time. As opposed to less-resourced languages like Persian, there are several text corpora in dominant languages like English which can be used for NLP applications. \\In this paper, MirasText which is an automatically generated text corpus for Persian language is presented. In this study, over 250 Persian websites were crawled and several fields like content, description, keywords, title, etc have been extracted to generate MirasText. Topic modeling and language modeling are used to validate the generated corpus. MirasText has over 2.8 million documents and over 1.4 billion tokens, which to our knowledge is the largest Persian corpus currently available.

The Reference Corpus of the Contemporary Romanian Language (CoRoLa)

Verginica Barbu Mititelu, Dan Tufiș and Elena Irimia

We present here the largest publicly available corpus of Romanian. Its written component contains 1,257,752,812 tokens, distributed, in an unbalanced way, in several language styles (legal, administrative, scientific, journalistic, imaginative, memoirs, blogposts), in four domains (arts and culture, nature, society, science) and in 71 subdomains. The oral component consists of almost 152 hours of recordings, with associated transcribed texts. All files have CMDI metadata associated. The written texts are automatically sentence-split, tokenized, part-of-speech tagged, lemmatized; a part of them are also syntactically annotated. The oral files are aligned with their corresponding transcriptions at word-phoneme level. The transcriptions are also automatically part-of-speech tagged, lemmatised and syllabified. CoRoLa contains original, IPR-cleared texts and is representative for the contemporary phase of the language, covering mostly the last 20 years. Its written component can be queried using the KorAP corpus management platform, whereas the oral component can be queried via its written counterpart, followed by the possibility of listening to the results of the query, using an in-house tool.

A Corpus of Drug Usage Guidelines Annotated with Type of Advice

Sarah Masud Preum, Md. Rizwan Parvez, Kai-Wei Chang and John Stankovic

Adherence to drug usage guidelines for prescription and over-the-counter drugs is critical for drug safety and effectiveness of treatment. Drug usage guideline documents contain advice on potential drug-drug interaction, drug-food interaction, and drug administration process. Current research on drug safety and public health indicates patients are often either unaware of such critical advice or overlook them. Categorizing advice statements from these documents according to their topics can enable the patients to find safety critical information. However, automatically categorizing drug usage guidelines based on their topic is an open challenge and there is no annotated dataset on drug usage guidelines. To address the latter issue, this paper presents (i) an annotation scheme for annotating safety critical advice from drug usage guidelines, (ii) an annotation tool for such data, and (iii) an annotated dataset containing drug usage guidelines from 90 drugs. This work is expected to accelerate further release of annotated drug usage guideline datasets and research on automatically filtering safety critical information from these textual documents.

BioRo: The Biomedical Corpus for the Romanian Language

Maria Mitrofan and Dan Tufis

The biomedical domain provides a large amount of linguistic resources usable for biomedical text mining. While most of the resources used in biomedical Natural Language Processing are available for English, for other languages including Romanian the access to language resources is not straight-forward. In this paper, we present the biomedical corpus of the Romanian language, which is a valuable linguistic asset for biomedical text mining.

This corpus was collected in the contexts of CoRoLa project, the reference corpus for the contemporary Romanian language. We also provide informative statistics about the corpus, a description of the data-composition. The annotation process of the corpus is also presented. Furthermore, we present the fraction of the corpus which will be made publicly available to the community without copyright restrictions.

Session P17 - Emotion Recognition/Generation

9th May 2018, 16:35

Chair person: **Lluís Padró** Poster Session

A Comparison Of Emotion Annotation Schemes And A New Annotated Data Set

Ian Wood, John Philip McCrae, Vladimir Andryushechkin and Paul Buitelaar

While the recognition of positive/negative sentiment in text is an established task with many standard data sets and well developed methodologies, the recognition of more nuanced affect has received less attention, and in particular, there are very few publicly available gold standard annotated resources. To address this lack, we present a series of emotion annotation studies on tweets culminating in a publicly available collection of 2,019 tweets with scores on four emotion dimensions: valence, arousal, dominance and surprise, following the emotion representation model identified by Fontaine et.al. (Fontaine et al., 2007). Further, we make a comparison of relative vs. absolute annotation schemes. We find improved annotator agreement with a relative annotation scheme (comparisons) on a dimensional emotion model over a categorical annotation scheme on Ekman's six basic emotions (Ekman et al., 1987), however when we compare inter-annotator agreement for comparisons with agreement for a rating scale annotation scheme (both with the same dimensional emotion model), we find improved inter-annotator agreement with rating scales, challenging a common belief that relative judgements are more reliable.

Humor Detection in English-Hindi Code-Mixed Social Media Content : Corpus and Baseline System

Ankush Khandelwal, Sahil Swami, Syed Sarfaraz Akhtar and Manish Shrivastava

The tremendous amount of user generated data through social networking sites led to the gaining popularity of automatic text classification in the field of computational linguistics over the past decade. Within this domain, one problem that has drawn the attention of many researchers is automatic humor detection in texts. In depth semantic understanding of the text is required to detect humor which makes the problem difficult to automate. With increase in the number of social media users, many multilingual speakers often interchange between languages while posting on social media which is called code-mixing. It introduces some challenges in the field of linguistic analysis of social media content (Barman et al., 2014), like spelling variations and non-grammatical structures in a sentence. Past researches include detecting puns in texts (Kao et al., 2016) and humor in one-lines (Mihalcea et al., 2010) in a single language, but with the tremendous amount of code-mixed data available online, there is a need to develop techniques which detects humor in code-mixed tweets. In this paper, we analyze the task of humor detection in texts and describe a freely available corpus containing English-Hindi code-mixed tweets annotated with humorous(H) or non-humorous(N) tags. We also tagged the words in the tweets with Language tags (English/Hindi/Others). Moreover, we provide a baseline classification system which distinguishes between humorous and non-humorous texts.

Dialogue Scenario Collection of Persuasive Dialogue with Emotional Expressions via Crowdsourcing

Koichiro Yoshino, Yoko Ishikawa, Masahiro Mizukami, Yu Suzuki, Sakriani Sakti and Satoshi Nakamura

Existing dialogue data collection methods such as the Wizard of Oz method (WoZ) or real dialogue recording are costly, and they prevent launching a new dialogue system. In this study, we requested crowd workers in crowdsourcing to create dialogue scenarios according to the instruction of the situation for persuasive dialogue systems that use emotional expressions. We collected 200 dialogues in 5 scenarios for a total of 1,000 via crowdsourcing. We also annotated emotional states and users' acceptance for system persuasion by using crowdsourcing. We constructed a persuasive dialogue system with the collected data and evaluated the system by interacting with crowd works. From the experiment, it was investigated that the collected labels have sufficient agreement even if we did not impose any training of annotation to workers.

SentiArabic: A Sentiment Analyzer for Standard Arabic

Ramy Eskander

Sentiment analysis has been receiving increasing interest as it conveys valuable information in regard to people's preferences and opinions. In this work, we present a sentiment analyzer that identifies the overall contextual polarity for Standard Arabic

text. The contribution of this work is threefold. First, we modify and extend SLSA; a large-scale Sentiment Lexicon for Standard Arabic. Second, we build a sentiment corpus of Standard Arabic text tagged for its contextual polarity. This corpus represents the training, development and test sets for the proposed system. Third, we build a lightweight lexicon-based sentiment analyzer for Standard Arabic (SentiArabic). The analyzer does not require running heavy computations, where the link to the lexicon is carried out through a morphological lookup as opposed to conducting a rich morphological analysis, while the assignment of the sentiment is based on a simple decision tree that uses polarity scores as opposed to a more complex machine learning approach that relies on lexical information, while negation receives special handling. The analyzer is highly efficient as it achieves an F-score of 76.5% when evaluated on a blind test set, which is the highest results reported for that set, and an absolute 3.0% increase over a state-of-the-art system that uses deep-learning models.

Contextual Dependencies in Time-Continuous Multidimensional Affect Recognition

Dmitrii Fedotov, Denis Ivanko, Maxim Sidorov and Wolfgang Minker

Modern research on emotion recognition often deals with time-continuously labelled spontaneous interactions. Such data is much closer to real world problems in contrast to utterance-level categorical labelling in acted emotion corpora that have widely been used to date. While working with time-continuous labelling, one usually uses context-aware models, such as recurrent neural networks. The amount of context needed to show the best performance should be defined in this case. Despite of the research done in this field there is still no agreement on this issue. In this paper we model different amounts of contextual input data by varying two parameters: sparsing coefficient and time window size. A series of experiments conducted with different modalities and emotional labels on the RECOLA corpora has shown a strong pattern between the amount of context used in model and performance. The pattern remains the same for different pairs of modalities and label dimensions, but the intensity differs. Knowledge about an appropriate context can significantly reduce the complexity of the model and increase its flexibility.

WikiArt Emotions: An Annotated Dataset of Emotions Evoked by Art

Saif Mohammad and Svetlana Kiritchenko

Art is imaginative human creation meant to be appreciated, make people think, and evoke an emotional response. Here for the first time, we create a dataset of more than 4,000 pieces of art (mostly paintings) that has annotations for emotions evoked in the observer. The pieces of art are selected from WikiArt.org's collection for four western styles (Renaissance Art, Post-Renaissance Art, Modern Art, and Contemporary Art). The art is annotated via crowdsourcing for one or more of twenty emotion categories (including neutral). In addition to emotions, the art is also annotated for whether it includes the depiction of a face and how much the observers like the art. The dataset, which we refer to as the {\it WikiArt Emotions Dataset}, can help answer several compelling questions, such as: what makes art evocative, how does art convey different emotions, what attributes of a painting make it well liked, what combinations of categories and emotions evoke strong emotional response, how much does the title of an art impact its emotional response, and what is the extent to which different categories of art evoke consistent emotions in people. We found that fear, happiness, love, and sadness were the dominant emotions that also obtained consistent annotations among the different annotators. We found that the title often impacts the affectual response to art. We show that pieces of art that depict faces draw more consistent emotional responses than those that do not. We also show, for each art category and emotion combination, the average agreements on the emotions evoked and the average art ratings. The WikiArt Emotions dataset also has applications in automatic image processing, as it can be used to develop systems that detect emotions evoked by art, and systems that can transform existing art (or even generate new art) that evokes the desired affectual response.

Arabic Data Science Toolkit: An API for Arabic Language Feature Extraction

Paul Rodrigues, Valerie Novak, C. Anton Rytting, Julie Yelle and Jennifer Boutz

We introduce Arabic Data Science Toolkit (ADST), a framework for Arabic language feature extraction, designed for data scientists that may not be familiar with Arabic or natural language processing. The functions in the toolkit allow data scientists to extend their algorithms beyond lexical or statistical methods and leverage Arabic-specific linguistic and stylistic features to enhance their systems and enable them to reach performance levels they might receive on languages with more resources, or languages with which they have more familiarity.

Sentence and Clause Level Emotion Annotation, Detection, and Classification in a Multi-Genre Corpus

Shabnam Tafreshi and Mona Diab

Predicting emotion categories (e.g. anger, joy, sadness) expressed by a sentence is challenging due to inherent multi-label smaller pieces such as phrases and clauses. To date, emotion has been studied in single genre, while models of human behaviors or situational awareness in the event of disasters require emotion

modeling in multi-genres. In this paper, we expand and unify existing annotated data in different genres (emotional blog post, news title, and movie reviews) using an inventory of 8 emotions from Plutchik's Wheel of Emotions tags. We develop systems for automatically detecting and classifying emotions in text, in different textual genres and granularity levels, namely, sentence and clause levels in a supervised setting. We explore the effectiveness of clause annotation in sentence-level emotion detection and classification (EDC). To our knowledge, our EDC system is the first to target the clause level; further we provide emotion annotation for movie reviews dataset for the first time.

Session P18 - Ethics and Legal Issues

9th May 2018, 16:35

Chair person: **Karën Fort** Poster Session

A Swedish Cookie-Theft Corpus

Dimitrios Kokkinakis, Kristina Lundholm Fors, Kathleen Fraser and Arto Nordlund

Language disturbances can be a diagnostic marker for neurodegenerative diseases, such as Alzheimer's disease, at earlier stages. Connected speech analysis provides a non-invasive and easy-to-assess measure for determining aspects of the severity of language impairment. In this paper we focus on the development of a new corpus consisting of audio recordings of picture descriptions (including transcriptions) of the Cookie-theft, produced by Swedish speakers. The speech elicitation procedure provides an established method of obtaining highly constrained samples of connected speech that can allow us to study the intricate interactions between various linguistic levels and cognition. We chose the Cookie-theft picture since it's a standardized test that has been used in various studies in the past, and therefore comparisons can be made based on previous research. This type of picture description task might be useful for detecting subtle language deficits in patients with subjective and mild cognitive impairment. The resulting corpus is a new, rich and multi-faceted resource for the investigation of linguistic characteristics of connected speech and a unique dataset that provides a rich resource for (future) research and experimentation in many areas, and of language impairment in particular. The information in the corpus can also be combined and correlated with other collected data about the speakers, such as neuropsychological tests, brain physiology and cerebrospinal fluid markers as well as imaging.

Sharing Copies of Synthetic Clinical Corpora without Physical Distribution — A Case Study to Get Around IPRs and Privacy Constraints Featuring the German JSYNCC Corpus

Christina Lohr, Sven Buechel and Udo Hahn

The legal culture in the European Union imposes almost unsurmountable hurdles to exploit copyright protected language data (in terms of intellectual property rights (IPRs) of media contents) and privacy protected medical health data (in terms of the notion of informational self-determination) as language resources for the NLP community. These juridical constraints have seriously hampered progress in resource-greedy NLP research, in particular for non-English languages in the clinical domain. In order to get around these restrictions, we introduce a novel approach for the creation and re-use of clinical corpora which is based on a two-step workflow. First, we substitute authentic clinical documents by synthetic ones, i.e., made-up reports and case studies written by medical professionals for educational purposes and published in medical e-textbooks. We thus eliminate patients' privacy concerns since no real, concrete individuals are addressed in such narratives. In a second step, we replace physical corpus distribution by sharing software for trustful re-construction of corpus copies. This is achieved by an end-to-end tool suite which extracts well-specified text fragments from e-books and assembles, on demand, identical copies of the same text corpus we defined at our lab at any other site where this software is executed. Thus, we avoid IPR violations since no physical corpus (raw text data) is distributed. As an illustrative case study which is easily portable to other languages we present JSYNCC, the largest and, even more importantly, first publicly available, corpus of German clinical language.

A Legal Perspective on Training Models for Natural Language Processing

Richard Eckart de Castilho, Giulia Dore, Thomas Margoni, Penny Labropoulou and Iryna Gurevych

A significant concern in processing natural language data is the often unclear legal status of the input and output data/resources. In this paper, we investigate this problem by discussing a typical activity in Natural Language Processing: the training of a machine learning model from an annotated corpus. We examine which legal rules apply at relevant steps and how they affect the legal status of the results, especially in terms of copyright and copyright-related rights.

Session P19 - LR Infrastructures and Architectures

9th May 2018, 16:35

Chair person: **Dieter van Uytvanck** Poster Session

LREMap, a Song of Resources and Evaluation

Riccardo Del Gratta, Sara Goggi, Gabriella Pardelli and Nicoletta Calzolari

After 8 years we revisit the LRE Map of Language Resources, introduced at LREC 2010, to try to get a picture of the field and its evolution as reflected by the creation and use of Language Resources. The purpose of the Map was in fact "to shed light on the vast amount of resources that represent the background of the research presented at LREC". It also aimed at a "change of culture in the field, actively engaging each researcher in the documentation task about resources". The data analysed here have been provided by the authors of several conferences during the phase of submission of papers, and contain information about ca. 7500 resources. We analysed the LRE Map data from many different viewpoints and the paper reports on the global picture, on different trends emerging from the diachronic perspective and finally on some comparisons between the 2 major conferences present in the Map: LREC and COLING.

Metadata Collection Records for Language Resources

Henk Van den Heuvel, Erwin Komen and Nelleke Oostdijk

In this paper we motivate the need for introducing more elaborate and consistent metadata records for collections of (linguistic) data resources in the CLARIN context. For this purpose we designed and implemented a CMDI profile. We validated the profile in a first pilot in which we populated the profile for 45 Dutch language resources. Given the complexity of the profile and special purpose requirements we developed our own interface for creating, editing, listing, copying and exporting descriptions of metadata collection records. The requirements for this interface and its implementation are described.

Managing Public Sector Data for Multilingual Applications Development

Stelios Piperidis, Penny Labropoulou, Miltos Deligiannis and Maria Giagkou

The current paper outlines the ELRC-SHARE repository, an infrastructure designed and developed in the framework of the European Language Resource Coordination action with the objective to host, document, manage and appropriately distribute language resources pertinent to machine translation, and specifically tailored to the needs of the eTranslation service of the European Commission. Due to the scope of the eTranslation service which seeks to facilitate multilingual communication across public administrations in 30 European countries and to enable Europe-wide multilingual digital services, ELRC-SHARE demonstrates a number of characteristics in terms of its technical and functional parameters, as well as in terms of its data management and documentation layers. The paper elaborates on the repository technical characteristics, the underlying metadata schema, the different ways in which data and metadata can be provided, the user roles and their respective permissions on data management, and, finally, the extensions currently being implemented.

Bridging the LAPPS Grid and CLARIN

Erhard Hinrichs, Nancy Ide, James Pustejovsky, Jan Hajic, Marie Hinrichs, Mohammad Fazleh Elahi, Keith Suderman, Marc Verhagen, Kyeongmin Rim, Pavel Stranak and Jozef Misutka

The LAPPS-CLARIN project is creating a "trust network" between the Language Applications (LAPPS) Grid and the WebLicht workflow engine hosted by the CLARIN-D Center in Tubingen. The project also includes integration of NLP services available from the LINDAT/CLARIN Center in Prague. The goal is to allow users on one side of the bridge to gain appropriately authenticated access to the other and enable seamless communication among tools and resources in both frameworks. The resulting "meta-framework" provides users across the globe with access to an unprecedented array of language processing facilities that cover multiple languages, tasks, and applications, all of which are fully interoperable.

Fluid Annotation: A Granularity-aware Annotation Tool for Chinese Word Fluidity

Shu-Kai HSIEH, Yu-Hsiang Tseng, Chi-Yao Lee and Chiung-Yu Chiang

This paper presents a novel word granularity-aware annotation framework for Chinese. Anchored in current functionalist linguistics, this model rearranges the boundary of word segmentation and linguistic annotation, and gears toward a deeper understanding of lexical units and their behavior. The web-based annotation UI also supports flexible annotation tasks for various linguistic and affective phenomena.

E-magyar – A Digital Language Processing System

Tamás Váradi, Eszter Simon, Bálint Sass, Iván Mittelholcz, Attila Novák, Balázs Indig, Richárd Farkas and Veronika Vincze

e-magyar is a new toolset for the analysis of Hungarian texts. It was produced as a collaborative effort of the Hungarian language technology community integrating the best state of the art tools, enhancing them where necessary, making them interoperable and releasing them with a clear license. It is a free, open, modular text processing pipeline which is integrated in the GATE system offering further prospects of interoperability. From tokenizing to parsing and named entity recognition, existing tools were examined and those selected for integration underwent various amount of overhaul in order to operate in the pipeline with a uniform encoding, and run in the same Java platform. The tokenizer was re-built from ground up and the flagship module, the morphological analyzer, based on the Humor system, was given a new annotation system and was implemented in the HFST framework. The system is aimed for a broad range of users, from language technology application developers to digital humanities researchers alike. It comes with a drag-and-drop demo on its website: http://e-magyar.hu/en/.

ILCM - A Virtual Research Infrastructure for Large-Scale Qualitative Data

Andreas Niekler, Arnim Bleier, Christian Kahmann, Lisa Posch, Gregor Wiedemann, Kenan Erdogan, Gerhard Heyer and Markus Strohmaier

The iLCM project pursues the development of an integrated research environment for the analysis of structured and unstructured data in a "Software as a Service" architecture (SaaS). The research environment addresses requirements for the quantitative evaluation of large amounts of qualitative data with text mining methods as well as requirements for the reproducibility of data-driven research designs in the social sciences. For this, the iLCM research environment comprises two central components. First, the Leipzig Corpus Miner (LCM), a decentralized SaaS application for the analysis of large amounts of news texts developed in a previous Digital Humanities project. Second, the text mining tools implemented in the LCM are extended by an "Open Research Computing" (ORC) environment for executable script documents, so-called "notebooks". This novel integration allows to combine generic, high-performance methods to process large amounts of unstructured text data and with individual program scripts to address specific research requirements in computational social science and digital humanities. ilcm.informatik.uni-leipzig.de

CLARIN's Key Resource Families

Darja Fišer, Jakob Lenardič and Tomaž Erjavec

CLARIN is a European Research Infrastructure that has been established to support the accessibility of language resources and technologies to researchers from the Digital Humanities and Social Sciences. This paper presents CLARIN's Key Resource Families, a new initiative within the infrastructure, the goal of which is to collect and present in a uniform way the most prominent data types in the network of CLARIN consortia that display a high degree of maturity, are available for most EU languages, are a rich source of social and cultural data, and as such are highly relevant for research from a wide range of disciplines and methodological approaches in the Digital Humanities and Social Sciences as well as for cross-disciplinary and trans-national comparative research. The four resource families that we present each in turn are newspaper, parliamentary, CMC (computer-mediated communication), and parallel corpora. We focus on their presentation within the infrastructure, their metadata in terms of size, temporal coverage, annotation, accessibility and license, and discuss current problems.

Indra: A Word Embedding and Semantic Relatedness Server

Juliano Efson Sales, Leonardo Souza, Siamak Barzegar, Brian Davis, André Freitas and Siegfried Handschuh

In recent years word embedding/distributional semantic models evolved to become a fundamental component in many natural language processing (NLP) architectures due to their ability of capturing and quantifying semantic associations at scale. Word embedding models can be used to satisfy recurrent tasks in NLP such as lexical and semantic generalisation in machine learning tasks, finding similar or related words and computing semantic relatedness of terms. However, building and consuming specific word embedding models require the setting of a large set of configurations, such as corpus-dependant parameters, distance measures as well as compositional models. Despite their increasing relevance as a component in NLP architectures, existing frameworks provide limited options in their ability to systematically build, parametrise, compare and evaluate different models. To answer this demand, this paper describes INDRA , a multi-lingual word embedding/distributional semantics framework which supports the creation, use and evaluation of word embedding models. In addition to the tool, INDRA also shares more than 65 pre-computed models in 14 languages.

A UIMA Database Interface for Managing NLP-related Text Annotations

Giuseppe Abrami and Alexander Mehler

NLP and automatic text analysis necessarily involve the annotation of natural language texts. The Apache Unstructured Information Management applications (UIMA) framework is used in several projects, tools and resources, and has become a de facto standard in this area. Despite the multiple use of UIMA as a document-based schema, it does not provide native database support. In order to facilitate distributed storage and enable UIMA-based projects to perform targeted queries, we have developed the UIMA Database Interface (UIMA DI). UIMA DI sets up an environment for a generic use of UIMA documents in database systems. In addition, the integration of UIMA DI into rights and resource management tools enables user and group-specific access to UIMA documents and provides data protection. Finally, UIMA documents can be made accessible for third party programs. UIMA DI, which we evaluate in relation to file system-based storage, is available under the GPLv3 license via GitHub.

European Language Resource Coordination: Collecting Language Resources for Public Sector Multilingual Information Management

Andrea Lösch, Valérie Mapelli, Stelios Piperidis, Andrejs Vasiljevs, Lilli Smal, Thierry Declerck, Eileen Schnur, Khalid Choukri and Josef Van Genabith

In order to help improve the quality, coverage and performance of automated translation solutions in the context of current and future Connecting Europe Facility (CEF) digital services, the European Language Resource Coordination (ELRC) consortium was set up through a service contract operating under the European Commission's CEF SMART 2014/1074 programme to initiate a number of actions to support the collection of Language Resources (LRs) within the public sector. The first action consisted in raising awareness in the public sector through the organisation of dedicated events: 2 conferences and 29 country-specific workshops to meet with national or regional/municipal governmental organisations, language competence centres, relevant European institutions and other potential holders of LRs from the public service administrations. In order to gather resources shared by the contributors, the ELRC-SHARE Repository was built up together with services to support the sharing of LRs, such as the ELRC Helpdesk and Intellectual property Rights (IPR) clearance support. All collected LRs should pass a validation process whose guidelines were developed within the project. The collected LRs cover all official EU languages, plus Icelandic and Norwegian.

Session I-O1: Industry Track - Industrial systems

10th May 2018, 09:45

Chair person: **Martin Jansche** Oral Session

Tilde MT Platform for Developing Client Specific MT Solutions

Mārcis Pinnis, Andrejs Vasiljevs, Rihards Kalniņš, Roberts Rozis, Raivis Skadiņš and Valters Šics

In this paper, we present Tilde MT, a custom machine translation platform that provides linguistic data storage (parallel, monolingual corpora, multilingual term collections), data cleaning and normalisation, statistical and neural machine translation system training and hosting functionality, as well as wide integration capabilities (a machine user API and popular computer-assisted translation tool plugins). We provide details for the most important features of the platform, as well as elaborate typical MT system training workflows for client-specific MT solution development.

Improving homograph disambiguation with supervised machine learning

Kyle Gorman, Gleb Mazovetskiy and Vitaly Nikolaev

We describe a pre-existing rule-based homograph disambiguation system used for text-to-speech synthesis at Google, and compare it against a novel system which performs disambiguation using classifiers trained on a small amount of labeled data. An evaluation of these systems, using a new, freely available English data set, finds that hybrid systems (making use of both rules and machine learning) are significantly more accurate than either hand-written rules or machine learning alone. The evaluation also finds minimal performance degradation when the hybrid system is configured to run on limited-resource mobile devices rather than on production servers. The two best systems described here are used for homograph disambiguation on all American English text-to-speech traffic at Google.

Text Normalization Infrastructure that Scales to Hundreds of Language Varieties

Mason Chua, Daan Van Esch, Noah Coccaro, Eunjoon Cho, Sujeet Bhandari and Libin Jia

We describe the automated multi-language text normalization infrastructure that prepares textual data to train language models used in Google's keyboards and speech recognition systems,

across hundreds of language varieties. Training corpora are sourced from various types of data sets, and the text is then normalized using a sequence of hand-written grammars and learned models. These systems need to scale to hundreds or thousands of language varieties in order to meet product needs. Frequent data refreshes, privacy considerations and simultaneous updates across such a high number of languages make manual inspection of the normalized training data infeasible, while there is ample opportunity for data normalization issues. By tracking metrics about the data and how it was processed, we are able to catch internal data processing issues and external data corruption issues that can be hard to notice using standard extrinsic evaluation methods. Showing the importance of paying attention to data normalization behavior in large-scale pipelines, these metrics have highlighted issues in Google's real-world speech recognition system that have caused significant, but latent, quality degradation.

Session O13 - Paraphrase & Semantics

10th May 2018, 09:45

Chair person: **Udo Kruschwitz** Oral Session

DeModify: A Dataset for Analyzing Contextual Constraints on Modifier Deletion

Vivi Nastase, Devon Fritz and Anette Frank

Tasks such as knowledge extraction, text simplification and summarization have in common the fact that from a text fragment a smaller (not necessarily contiguous) portion is obtained by discarding part of the context. This may cause the text fragment to acquire a new meaning, or even to become false. The smallest units that can be considered disposable in a larger context are modifiers. In this paper we describe a dataset collected and annotated to facilitate the study of the influence of modifiers on the meaning of the context they are part of, and to support the development of models that can determine whether a modifier can be removed without undesirable semantic consequences.

Open Subtitles Paraphrase Corpus for Six Languages

Mathias Creutz

This paper accompanies the release of Opusparcus, a new paraphrase corpus for six European languages: German, English, Finnish, French, Russian, and Swedish. The corpus consists of paraphrases, that is, pairs of sentences in the same language that mean approximately the same thing. The paraphrases are extracted from the OpenSubtitles2016 corpus, which contains subtitles from movies and TV shows. The informal and colloquial genre that occurs in subtitles makes such data a very interesting language resource, for instance, from the perspective of computer assisted language learning. For each target language, the Opusparcus data have been partitioned into three types of data sets: training, development and test sets. The training sets are large, consisting of millions of sentence pairs, and have been compiled automatically, with the help of probabilistic ranking functions. The development and test sets consist of sentence pairs that have been checked manually; each set contains approximately 1000 sentence pairs that have been verified to be acceptable paraphrases by two annotators.

Fine-grained Semantic Textual Similarity for Serbian

Vuk Batanović, Miloš Cvetanović and Boško Nikolić

Although the task of semantic textual similarity (STS) has gained in prominence in the last few years, annotated STS datasets for model training and evaluation, particularly those with fine-grained similarity scores, remain scarce for languages other than English, and practically non-existent for minor ones. In this paper, we present the Serbian Semantic Textual Similarity News Corpus (STS.news.sr) – an STS dataset for Serbian that contains 1192 sentence pairs annotated with fine-grained semantic similarity scores. We describe the process of its creation and annotation, and we analyze and compare our corpus with the existing news-based STS datasets in English and other major languages. Several existing STS models are evaluated on the Serbian STS News Corpus, and a new supervised bag-of-words model that combines part-of-speech weighting with term frequency weighting is proposed and shown to outperform similar methods. Since Serbian is a morphologically rich language, the effect of various morphological normalization tools on STS model performances is considered as well. The Serbian STS News Corpus, the annotation tool and guidelines used in its creation, and the STS model framework used in the evaluation are all made publicly available.

SPADE: Evaluation Dataset for Monolingual Phrase Alignment

Yuki Arase and Jun'ichi Tsujii

We create the SPADE (Syntactic Phrase Alignment Dataset for Evaluation) for systematic research on syntactic phrase alignment in paraphrasal sentences. This is the first dataset to shed lights on syntactic and phrasal paraphrases under linguistically motivated grammar. Existing datasets available for evaluation on phrasal paraphrase detection define the unit of phrase as simply sequence of words without syntactic structures due to difficulties caused by the non-homographic nature of phrase correspondences in

sentential paraphrases. Different from these, the SPADE provides annotations of gold parse trees by a linguistic expert and gold phrase alignments identified by three annotators. Consequently, 20,276 phrases are extracted from 201 sentential paraphrases, on which 15,721 alignments are obtained that at least one annotator regarded as paraphrases. The SPADE is available at Linguistic Data Consortium for future research on paraphrases. In addition, two metrics are proposed to evaluate to what extent the automatic phrase alignment results agree with the ones identified by humans. These metrics allow objective comparison of performances of different methods evaluated on the SPADE. Benchmarks to show performances of humans and the state-of-the-art method are presented as a reference for future SPADE users.

ETPC - A Paraphrase Identification Corpus Annotated with Extended Paraphrase Typology and Negation

Venelin Kovatchev, Toni Marti and Maria Salamo

We present the Extended Paraphrase typology (EPT) and the Extended Typology Paraphrase Corpus (ETPC). The EPT typology addresses several practical limitations of existing paraphrase typologies: it is the first typology that copes with the non-paraphrase pairs in the paraphrase identification corpora and distinguishes between contextual and habitual paraphrase types. ETPC is the largest corpus to date annotated with atomic paraphrase types. It is the first corpus with detailed annotation of both the paraphrase and the non-paraphrase pairs and the first corpus annotated with paraphrase and negation. Both new resources contribute to better understanding the paraphrase phenomenon, and allow for studying the relationship between paraphrasing and negation. To the developers of Paraphrase Identification systems ETPC corpus offers better means for evaluation and error analysis. Furthermore, the EPT typology and ETPC corpus emphasize the relationship with other areas of NLP such as Semantic Similarity, Textual Entailment, Summarization and Simplification.

Session O14 - Emotion & Sentiment (2)

10th May 2018, 09:45

Chair person: **Min Zhang** Oral Session

Introducing a Lexicon of Verbal Polarity Shifters for English

Marc Schulder, Michael Wiegand, Josef Ruppenhofer and Stephanie Köser

The sentiment polarity of a phrase does not only depend on the polarities of its words, but also on how these are affected by their context. Negation words (e.g. "not", "no", "never") can change the polarity of a phrase. Similarly, verbs and other content words can also act as polarity shifters (e.g. "fail", "deny", "alleviate"). While individually more sparse, they are far more numerous. Among verbs alone, there are more than 1200 shifters. However, sentiment analysis systems barely consider polarity shifters other than negation words. A major reason for this is the scarcity of lexicons and corpora that provide information on them. We introduce a lexicon of verbal polarity shifters that covers the entirety of verbs found in WordNet. We provide a fine-grained annotation of individual word senses, as well as information for each verbal shifter on the syntactic scopes that it can affect.

JFCKB: Japanese Feature Change Knowledge Base

Tetsuaki Nakamura and Daisuke Kawahara

Commonsense knowledge plays an essential role in our language activities. Although many projects have aimed to develop language resources for commonsense knowledge, there is little work focusing on connotational meanings. This is because constructing commonsense knowledge including connotational meanings is challenging. In this paper, we present a Japanese knowledge base where arguments in event sentences are associated with various feature changes caused by the events. For example, “my child” in “my wife hits my child” is associated with some feature changes, such as increase in pain, increase in anger, increase in disgust, and decrease in joy. We constructed this knowledge base through crowdsourcing tasks by gathering feature changes of arguments in event sentences. After the construction of the knowledge base, we conducted an experiment in anaphora resolution using the knowledge base. We regarded anaphora resolution as an antecedent candidate ranking task and used Ranking SVM as the solver. Experimental results demonstrated the usefulness of our feature change knowledge base.

Quantifying Qualitative Data for Understanding Controversial Issues

Michael Wojatzki, Saif Mohammad, Torsten Zesch and Svetlana Kiritchenko

Understanding public opinion on complex controversial issues such as ‘Legalization of Marijuana’ is of considerable importance for a number of objectives. However, an individual’s position on a controversial issue is often not a binary support-or-oppose stance on the issue, but rather a conglomerate of nuanced opinions on various aspects of the issue. These opinions are often expressed qualitatively in free text in surveys or on social media. However, quantifying vast amounts of qualitative information remains a significant challenge. The goal of this work is to provide a new

approach for quantifying qualitative data for the understanding of controversial issues. First, we show how we can engage people directly through crowdsourcing to create a comprehensive dataset of assertions (claims, opinions, etc.) relevant to an issue. Next, the assertions are judged for agreement and strength of support or opposition. The collected Dataset of Nuanced Assertions on Controversial Issues (NAoCI dataset) consists of over 2,000 assertions on sixteen different controversial issues. It has over 100,000 judgments of whether people agree or disagree with the assertions, and of about 70,000 judgments indicating how strongly people support or oppose the assertions. This dataset allows for several useful analyses that help summarize public opinion.

Distribution of Emotional Reactions to News Articles in Twitter

Omar Juárez Gambino, Hiram Calvo and Consuelo-Varinia García-Mendoza

Several datasets of opinions expressed by Social networks' users have been created to explore Sentiment Analysis tasks like Sentiment Polarity and Emotion Mining. Most of these datasets are focused on the writers' perspective, that is, the post written by a user is analyzed to determine the expressed sentiment on it. This kind of datasets do not consider the source that provokes those opinions (e.g. a previous post). In this work, we propose a dataset focused on the readers' perspective. The developed dataset contains news articles published by three newspapers and the distribution of six predefined emotions expressed by readers of the articles in Twitter. This dataset was built aiming to explore how the six emotions are expressed by Twitter users' after reading a news article. We show some results of a machine learning method used to predict the distribution of emotions in unseen news articles.

Aggression-annotated Corpus of Hindi-English Code-mixed Data

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia and Tushar Maheshwari

As the interaction over the web has increased, incidents of aggression and related events like trolling, cyberbullying, flaming, hate speech, etc. too have increased manifold across the globe. While most of these behaviour like bullying or hate speech have predated the Internet, the reach and extent of the Internet has given these an unprecedented power and influence to affect the lives of billions of people. So it is of utmost significance and importance that some preventive measures be taken to provide safeguard to the people using the web such that the web remains a viable medium of communication and connection, in general. In this paper, we discuss the development of an aggression tagset and an annotated corpus of Hindi-English code-mixed data from two of the most popular social networking / social media platforms in India – Twitter and Facebook. The corpus is annotated using a hierarchical tagset of 3 top-level tags and 10 level 2 tags. The final dataset contains approximately 18k tweets and 21k facebook comments and is being released for further research in the field.

Session O15 - Semantics & Lexicon (2)

10th May 2018, 09:45

Chair person: **Reinhard Rapp** Oral Session

Creating a Verb Synonym Lexicon Based on a Parallel Corpus

Zdenka Uresova, Eva Fucikova, Eva Hajicova and Jan Hajic

This paper presents the first findings of our recently started project of building a new lexical resource called CzEngClass, which consists of bilingual verbal synonym groups. In order to create such a resource, we explore semantic 'equivalence' of verb senses (across different verb lexemes) in a bilingual (Czech-English) setting by using translational context of real-world texts in a parallel, richly annotated dependency corpus. When grouping semantically equivalent verb senses into classes of synonyms, we focus on valency (arguments as deep dependents with morphosyntactic features relevant for surface dependencies) and its mapping to a set of semantic "roles" for verb arguments, common within one class. We argue that the existence of core argument mappings and certain adjunct mappings to a common set of semantic roles is a suitable criterion for a reasonable verb synonymy definition, possibly accompanied with additional contextual restrictions. By mid-2018, the first version of the lexicon called CzEngClass will be publicly available.

Evaluation of Domain-specific Word Embeddings using Knowledge Resources

Farhad Nooralahzadeh, Lilja Øvrelid and Jan Tore Lønning

In this work we evaluate domain-specific embedding models induced from textual resources in the Oil and Gas domain. We conduct intrinsic and extrinsic evaluations of both general and domain-specific embeddings and we observe that constructing domain-specific word embeddings is worthwhile even with a considerably smaller corpus size. Although the intrinsic evaluation shows low performance in synonymy detection, an in-depth error analysis reveals the ability of these models to discover additional semantic relations such as hyponymy, co-hyponymy and relatedness in the target domain. Extrinsic evaluation of

the embedding models is provided by a domain-specific sentence classification task, which we solve using a convolutional neural network. We further adapt embedding enhancement methods to provide vector representations for infrequent and unseen terms. Experiments show that the adapted technique can provide improvements both in intrinsic and extrinsic evaluation.

Automatic Thesaurus Construction for Modern Hebrew

Chaya Liebeskind, Ido Dagan and Jonathan Schler

Automatic thesaurus construction for Modern Hebrew is a complicated task, due to its high degree of inflectional ambiguity. Linguistics tools, including morphological analyzers, part-of-speech taggers and parsers often have limited in performance on Morphologically Rich Languages (MRLs) such as Hebrew. In this paper, we adopted a schematic methodology for generating a co-occurrence based thesaurus in a MRL and extended the methodology to create distributional similarity thesaurus. We explored three alternative levels of morphological term representations, surface form, lemma, and multiple lemmas, all complemented by the clustering of morphological variants. First, we evaluated both the co-occurrence based method and the distributional similarity method using Hebrew WordNet as our gold standard. However, due to Hebrew WordNet's low coverage, we completed our analysis with a manual evaluation. The results showed that for Modern Hebrew corpus-based thesaurus construction, the most directly applied statistical collection, using linguistics tools at the lemma level, is not optimal.

Automatic Wordnet Mapping: from CoreNet to Princeton WordNet

Jiseong Kim, Younggyun Hahm, Sunggoo Kwon and KEY-SUN CHOI

CoreNet is a lexico-semantic network of 73,100 Korean word senses, which are categorized under 2,937 semantic categories organized in a taxonomy. Recently, to foster the more widespread use of CoreNet, there was an attempt to map the semantic categories of CoreNet into synsets of Princeton WordNet by lexical relations such as synonymy, hyponymy, and hypernymy relations. One of the limitations of the existing mapping is that it is only focused on mapping the semantic categories, but not on mapping the word senses, which are the majority part (96%) of CoreNet. To boost bridging the gap between CoreNet and WordNet, we introduce the automatic mapping approach to link the word senses of CoreNet into WordNet synsets. The evaluation shows that our approach successfully maps previously unmapped 38,028 word senses into WordNet synsets with the precision of 91.2% (ś1.14 with 99% confidence).

The New Propbank: Aligning Propbank with AMR through POS Unification

Tim O'Gorman, Sameer Pradhan, Martha Palmer, Julia Bonn, Kathryn Conger and James Gung

We present a corpus which converts the sense labels of existing Propbank resources to a new unified format which is more compatible with AMR and more robust to sparsity. This adopts an innovation of the Abstract Meaning Representation project(Banarescu et al. 2013) in which one abstracts away from different, related parts of speech, so that related forms such as "insert" and "insertion" could be represented by the same roleset and use the same semantic roles. We note that this conversion also serves to make the different English Propbank corpora released over the years consistent with each other, so that one might train and evaluate systems upon that larger combined data. We present analysis of some appealing characteristics of this final dataset, and present preliminary results of training and evaluating SRL systems on this combined set, to spur usage of this challenging new dataset.

Session O16 - Bilingual Speech Corpora & Code-Switching

10th May 2018, 09:45

Chair person: **Christopher Cieri** Oral Session

The Boarnsterhim Corpus: A Bilingual Frisian-Dutch Panel and Trend Study

Marjoleine Sloos, Eduard Drenth and Wilbert Heeringa

The Boarnsterhim Corpus consists of 250 hours of speech in both West Frisian and Dutch by the same sample of bilingual speakers. The corpus contains original recordings from 1982-1984 and a replication study recorded 35 years later. The data collection spans speech of four generations, and combines panel and trend data. This paper describes the Boarnsterhim Corpus halfway the project which started in 2016 and describes the way it was collected, the annotations, potential use, and the envisaged tools and end-user web application.

The French-Algerian Code-Switching Triggered audio corpus (FACST)

Amazouz Djegdjiga, Martine Adda-Decker and Lori Lamel

The French Algerian Code-Switching Triggered corpus (FACST) was created in order to support a variety of studies in phonetics, prosody and natural language processing. The first aim of the FACST corpus is to collect a spontaneous Code-switching speech (CS) corpus. In order to obtain a large quantity of spontaneous CS utterances in natural conversations experiments were carried

out on how to elicit CS. Applying a triggering protocol by means of code-switched questions was found to be effective in eliciting CS in the responses. To ensure good audio quality, all recordings were made in a soundproof room or in a very calm room. This paper describes FACST corpus, along with the principal steps to build a CS speech corpus in French-Algerian languages and data collection steps. We also explain the selection criteria for the CS speakers and the recording protocols used. We present the methods used for data segmentation and annotation, and propose a conventional transcription of this type of speech in each language with the aim of being well-suited for both computational linguistic and acoustic-phonetic studies. We provide an a quantitative description of the FACST corpus along with results of linguistic studies, and discuss some of the challenges we faced in collecting CS data.

Strategies and Challenges for Crowdsourcing Regional Dialect Perception Data for Swiss German and Swiss French

Jean-Philippe Goldman, Simon Clematide, Mathieu Avanzi and Raphaël Tandler

Following the dynamics of several recent crowdsourcing projects with the aim of collecting linguistic data, this paper focuses on such a project in the field of Swiss German dialects and Swiss French accents. The main scientific goal of the data collected is to understand people's perception of dialects and accents, and provide a resource for future computational systems such as automatic dialect recognition. A gamified crowdsourcing platform was set up and launched for both main locales of Switzerland: "din dialäkt" ('your dialect') for Swiss German dialects and "ton accent" ('your accent') for Swiss French. The main activity for the participant is to localize preselected audio samples by clicking on a map of Switzerland. The media was highly interested in the two platforms and many reports appeared in newspapers, television and radio, which increased the public's awareness of the project and thus also the traffic on the page. At this point of the project, 7,500 registered users (beside 30,000 anonymous visitors), have provided 470,000 localizations. By connecting user's results of this localization task to their socio-demographic information, a quantitative analysis of the localization data can reveal which factors play a role in their performance. Preliminary results showed that age and childhood residence influence the how well dialects/accents are recognized. Nevertheless, quantity does not ensure quality when it comes to data. Crowdsourcing such linguistic data revealed traps to avoid such as scammers, or the participants' quick loss of motivation causing them to click randomly. Such obstacles need to be taken into account when assessing the reliability of data and require a number of preliminary steps before an analysis of the data.

Phonetically Balanced Code-Mixed Speech Corpus for Hindi-English Automatic Speech Recognition

Ayushi Pandey, Brij Mohan Lal Srivastava, Rohit Kumar, Bhanu Teja Nellore, Kasi Sai Teja and Suryakanth V Gangashetty

The paper presents the development of a phonetically balanced read speech corpus of code-mixed Hindi-English. Phonetic balance in the corpus has been created by selecting sentences that contained triphones lower in frequency than a predefined threshold. The assumption with a compulsory inclusion of such rare units was that the high frequency triphones will inevitably be included. Using this metric, the Pearson's correlation coefficient of the phonetically balanced corpus with a large code-mixed reference corpus was recorded to be 0.996. The data for corpus creation has been extracted from selected sections of Hindi newspapers.These sections contain frequent English insertions in a matrix of Hindi sentence. Statistics on the phone and triphone distribution have been presented, to graphically display the phonetic likeness between the reference corpus and the corpus sampled through our method.

Chinese-Portuguese Machine Translation: A Study on Building Parallel Corpora from Comparable Texts

Siyou Liu, Longyue Wang and Chao-Hong Liu

Although there are increasing and significant ties between China and Portuguese-speaking countries, there is not much parallel corpora in the Chinese-Portuguese language pair. Both languages are very populous, with 1.2 billion native Chinese speakers and 279 million native Portuguese speakers, the language pair, however, could be considered as low-resource in terms of available parallel corpora. In this paper, we describe our methods to curate Chinese-Portuguese parallel corpora and evaluate their quality. We extracted bilingual data from Macao government websites and proposed a hierarchical strategy to build a large parallel corpus. Experiments are conducted on existing and our corpora using both Phrased-Based Machine Translation (PBMT) and the state-of-the-art Neural Machine Translation (NMT) models. The results of this work can be used as a benchmark for future Chinese-Portuguese MT systems. The approach we used in this paper also show a good example on how to boost performance of MT systems for low-resource language pairs.

Session P20 - Bibliometrics, Scientometrics, Infometrics

10th May 2018, 09:45

Chair person: **Richard Eckart de Castilho** Poster Session

A High-Quality Gold Standard for Citation-based Tasks

Michael Färber, Alexander Thiemann and Adam Jatowt

Analyzing and recommending citations within their specific citation contexts has recently received much attention due to the growing number of available publications. Although data sets such as CiteSeerX have been created for evaluating approaches for such tasks, those data sets exhibit striking defects. This is understandable when one considers that both information extraction and entity linking, as well as entity resolution, need to be performed. In this paper, we propose a new evaluation data set for citation-dependent tasks based on arXiv.org publications. Our data set is characterized by the fact that it exhibits almost zero noise in its extracted content and that all citations are linked to their correct publications. Besides the pure content, available on a sentence-by-sentence basis, cited publications are annotated directly in the text via global identifiers. As far as possible, referenced publications are further linked to the DBLP Computer Science Bibliography. Our data set consists of over 15 million sentences and is freely available for research purposes. It can be used for training and testing citation-based tasks, such as recommending citations, determining the functions or importance of citations, and summarizing documents based on their citations.

Measuring Innovation in Speech and Language Processing Publications.

Joseph Mariani, Gil Francopoulo and Patrick Paroubek

The goal of this paper is to propose measures of innovation through the study of publications in the field of speech and language processing. It is based on the NLP4NLP corpus, which contains the articles published in major conferences and journals related to speech and language processing over 50 years (1965-2015). It represents 65,003 documents from 34 different sources, conferences and journals, published by 48,894 different authors in 558 events, for a total of more than 270 million words and 324,422 bibliographical references. The data was obtained in textual form or as an image that had to be converted into text. This resulted in a lower quality for the most ancient papers, that we measured through the computation of an unknown word ratio. The multi-word technical terms were automatically extracted after parsing, using a set of general language text corpora. The occurrences, frequencies, existences and presences of the terms were then computed overall, for each year and for each document. It resulted in a list of 3.5 million different terms and 24 million term occurrences. The evolution of the research topics over the year, as reflected by the terms presence, was then computed and we propose a measure of the topic popularity based on this computation. The author(s) who introduced the terms were searched for, together with the year when the term was first introduced and the publication where it was introduced. We then studied the global and evolutional contributions of authors to a given topic. We also studied the global and evolutional contributions of the various publications to a given topic. We finally propose a measure of innovativeness for authors and publications.

PDFdigest: an Adaptable Layout-Aware PDF-to-XML Textual Content Extractor for Scientific Articles

Daniel Ferrés, Horacio Saggion, Francesco Ronzano and Àlex Bravo

The availability of automated approaches and tools to extract structured textual content from PDF articles is essential to enable scientific text mining. This paper describes and evaluates the PDFdigest tool, a PDF-to-XML textual content extraction system specially designed to extract scientific articles' headings and logical structure (title, authors, abstract,...) and its textual content. The extractor deals with both text-based and image-based PDF articles using custom rule-based algorithms based on existing state-of-the-art open-source tools for both PDF-to-HTML conversion and image-based PDF Optical Character Recognition.

Automatic Identification of Research Fields in Scientific Papers

Eric Kergosien, Amin Farvardin, Maguelonne Teisseire, Marie-Noelle BESSAGNET, Joachim Schöpfel, Stéphane Chaudiron, Bernard Jacquemin, Annig Lacayrelle, Mathieu Roche, Christian Sallaberry and Jean-Philippe Tonneau

The TERRE-ISTEX project aims to identify scientific research dealing with specific geographical territories areas based on heterogeneous digital content available in scientific papers. The project is divided into three main work packages: (1) identification of the periods and places of empirical studies, and which reflect the publications resulting from the analyzed text samples, (2) identification of the themes which appear in these documents, and (3) development of a web-based geographical information retrieval tool (GIR). The first two actions combine Natural Language Processing patterns with text mining methods. The integration of the spatial, thematic and temporal dimensions in a GIR contributes to a better understanding of what kind of research has been carried out, of its topics and its geographical and historical coverage. Another originality of the TERRE-ISTEX

project is the heterogeneous character of the corpus, including PhD theses and scientific articles from the ISTEX digital libraries and the CIRAD research center.

Session P21 - Discourse Annotation, Representation and Processing (1)

10th May 2018, 09:45

Chair person: **Silvia Pareti** Poster Session

A «Portrait» Approach to Multichannel Discourse

Andrej Kibrik and Olga Fedorova

This paper contributes to the research field of multichannel discourse analysis. Multimodal discourse analysis explores numerous channels involved in natural communication, such as verbal structure, prosody, gesticulation, facial expression, eye gaze, etc., and treats them as parts of an integral process. Among the key issues in multichannel studies is the question of the individual variation in multichannel behavior. We address this issue with the help of a multichannel resource "Russian Pear Chats and Stories" that is currently under construction (multidiscourse.ru). This corpus is based on a novel methodology of data collection and is produced with the help of state of the art technology including eyetracking. To address the issue of individual variation, we introduce the notion of a speaker's individual portrait. In particular, we consider the Prosodic Portrait, the Oculomotor Portrait, and the Gesticulation Portrait. The proposed methodology is crucially important for fine-grained annotation procedures as well as for accurate statistic analyses of multichannel data.

Multilingual Extension of PDTB-Style Annotation: The Case of TED Multilingual Discourse Bank

Deniz Zeyrek, Amália Mendes and Murathan Kurfalı

We introduce TED-Multilingual Discourse Bank, a corpus of TED talks transcripts in 6 languages (English, German, Polish, European Portuguese, Russian and Turkish), where the ultimate aim is to provide a clearly described level of discourse structure and semantics in multiple languages. The corpus is manually annotated following the goals and principles of PDTB, involving explicit and implicit discourse connectives, entity relations, alternative lexicalizations and no relations. In the corpus, we also aim to capture the characteristics of spoken language that exist in the transcripts and adapt the PDTB scheme according to our aims; for example, we introduce hypophora. We spot other aspects of spoken discourse such as the discourse marker use of connectives to keep them distinct from their discourse connective use. TED-MDB is, to the best of our knowledge, one of the few multilingual discourse treebanks and is hoped to be a source of parallel data for contrastive linguistic analysis as well as language technology applications. We describe the corpus, the annotation procedure and provide preliminary corpus statistics.

Building a Macro Chinese Discourse Treebank

Xiaomin Chu, Feng Jiang, Sheng Xu and Qiaoming Zhu

Discourse structure analysis is an important research topic in natural language processing. Discourse structure analysis not only helps to understand the discourse structure and semantics, but also provides strong support for deep applications of natural language processing, such as automatic summarization, statistical machine translation, question and answering, etc. At present, the analyses of discourse structure are mainly concentrated on the micro level, while the analyses on macro level are few. Therefore, this paper focuses on the construction of representation schema and corpus resources on the macro level of discourse structure. This paper puts forward a macro discourse structure framework and constructs the logical semantic structure and functional pragmatic structure respectively. On this basis, a macro Chinese discourse structure treebank is annotated, consisting of 147 Newswire articles. Preliminary experimental results show that the representation schema and corpus resource constructed in this paper can lay the foundation for further analysis of macro discourse structure.

Enhancing the AI2 Diagrams Dataset Using Rhetorical Structure Theory

Tuomo Hiippala and Serafina Orekhova

This paper describes ongoing work on a multimodal resource based on the Allen Institute AI2 Diagrams (AI2D) dataset, which contains nearly 5000 grade-school level science diagrams that have been annotated for their elements and the semantic relations that hold between them. This emerging resource, named AI2D-RST, aims to provide a drop-in replacement for the annotation of semantic relations between diagram elements, whose description is informed by recent theories of multimodality and text-image relations. As the name of the resource suggests, the revised annotation schema is based on Rhetorical Structure Theory (RST), which has been previously used to describe the multimodal structure of diagrams and entire documents. The paper documents the proposed annotation schema, describes challenges in applying RST to diagrams, and reports on inter-annotator agreement for this task. Finally, the paper discusses the use of AI2D-RST for research on multimodality and artificial intelligence.

QUD-Based Annotation of Discourse Structure and Information Structure: Tool and Evaluation

Kordula De Kuthy, Nils Reiter and Arndt Riester

We discuss and evaluate a new annotation scheme and discourse-analytic method, the QUD-tree framework. We present an annotation study, in which the framework, based on the concept of Questions under Discussion, is applied to English and German interview data, using TreeAnno, an annotation tool specially developed for this new kind of discourse annotation. The results of an inter-annotator agreement study show that the new annotation method allows for reasonable agreement with regard to discourse structure and good agreement with regard to the annotation of information structure, which covers focus, background, contrastive topic and non-at-issue material.

The Spot the Difference corpus: a multi-modal corpus of spontaneous task oriented spoken interactions

José Lopes, Nils Hemmingsson and Oliver Åstrand

This paper describes the Spot the Difference Corpus which contains 54 interactions between pairs of subjects interacting to find differences in two very similar scenes. The setup used, the participants' metadata and details about collection are described. We are releasing this corpus of task-oriented spontaneous dialogues. This release includes rich transcriptions, annotations, audio and video. We believe that this dataset constitutes a valuable resource to study several dimensions of human communication that go from turn-taking to the study of referring expressions. In our preliminary analyses we have looked at task success (how many differences were found out of the total number of differences) and how it evolves over time. In addition we have looked at scene complexity provided by the RGB components' entropy and how it could relate to speech overlaps, interruptions and the expression of uncertainty. We found there is a tendency that more complex scenes have more competitive interruptions.

Attention for Implicit Discourse Relation Recognition

Andre Cianflone and Leila Kosseim

Implicit discourse relation recognition remains a challenging task as state-of-the-art approaches reach F1 scores ranging from 9.95% to 37.67% on the 2016 CoNLL shared task. In our work, we explore the use of a neural network which exploits the strong correlation between pairs of words across two discourse arguments that implicitly signal a discourse relation. We present a novel approach to Implicit Discourse Relation Recognition that uses an encoder-decoder model with attention. Our approach is based on the assumption that a discourse argument is "generated" from a previous argument and conditioned on a latent discourse relation, which we detect. Experiments show that our model achieves an F1 score of 38.25% on fine-grained classification, outperforming previous approaches and performing comparatively with state-of-the-art on coarse-grained classification, while computing alignment parameters without the need for additional pooling and fully connected layers.

A Context-based Approach for Dialogue Act Recognition using Simple Recurrent Neural Networks

Chandrakant Bothe, Cornelius Weber, Sven Magg and Stefan Wermter

Dialogue act recognition is an important part of natural language understanding. We investigate the way dialogue act corpora are annotated and the learning approaches used so far. We find that the dialogue act is context-sensitive within the conversation for most of the classes. Nevertheless, previous models of dialogue act classification work on the utterance-level and only very few consider context. We propose a novel context-based learning method to classify dialogue acts using a character-level language model utterance representation, and we notice significant improvement. We evaluate this method on the Switchboard Dialogue Act corpus, and our results show that the consideration of the preceding utterances as a context of the current utterance improves dialogue act detection.

TreeAnnotator: Versatile Visual Annotation of Hierarchical Text Relations

Philipp Helfrich, Elias Rieb, Giuseppe Abrami, Andy Lücking and Alexander Mehler

We introduce TreeAnnotator, a graphical tool for annotating tree-like structures, in particular structures that jointly map dependency relations and inclusion hierarchies, as used by Rhetorical Structure Theory (RST). TreeAnnotator is browser-based, embedded within the UIMA framework and provides two visualization modes. TreeAnnotator's interoperability exceeds similar tools, providing a wider range of formats, while annotation work can be completed more quickly due to a revised input method for RST dependency relations. TreeAnnotator offers a multiple window view, which allows users to inspect several annotations side by side. For storing and versioning annotations, the UIMA Database Interface (UIMA DI) was developed to save documents based on a pre-defined type system. These features not only connect TreeAnnotator annotations to modern technological and dialog theoretical work, but set it apart from related tools. The ease of use of TreeAnnotator and its newly designed user interface is evaluated in a user study consisting of annotating rhetorical relations with TreeAnnotator and the classic RSTTool.

Chats and Chunks: Annotation and Analysis of Multiparty Long Casual Conversations

Emer Gilmartin, Carl Vogel and Nick Campbell

Casual talk or social conversation is a fundamental form of spoken interaction. Corpora of casual talk often comprise relatively short dyadic conversations, although research into such talk has found longer multiparty interaction to be very common. This genre of spoken interaction is attracting more interest with attempts to build more friendly and natural spoken dialog systems. To study longer multiparty casual talk, we have assembled a collection of conversations from three existing corpora. We describe the collection, organization, and annotation of structural chat and chunk phases in these conversations. we then review our preliminary results, noting significant differences in the distribution of overlap, laughter and disfluency in chat and chunk phases, and finding that chunk dominates as conversations get longer. We outline our continuing work on gaining greater understanding of this genre of spoken interaction, with implications for the design of spoken dialog systems.

Session P22 - Evaluation Methodologies

10th May 2018, 09:45

Chair person: **Edouard Geoffrois** Poster Session

Extending the gold standard for a lexical substitution task: is it worth it?

Ludovic Tanguy, Cécile Fabre and Laura Rivière

We present a new evaluation scheme for the lexical substitution task. Following (McCarthy and Navigli, 2007) we conducted an annotation task for French that mixes two datasets: in the first one, 300 sentences containing a target word (among 30 different) were submitted to annotators who were asked to provide substitutes. The second one contains the propositions of the systems that participated to the lexical substitution task based on the same data. The idea is first, to assess the capacity of the systems to provide good substitutes that would not have been proposed by the annotators and second, to measure the impact on the task evaluation of a new gold standard that incorporates these additional data. While (McCarthy and Navigli, 2009) have conducted a similar post hoc analysis, re-evaluation of the systems' performances has not been carried out to our knowledge. This experiment shows interesting differences between the two resulting datasets and gives insight on how automatically retrieved substitutes can provide complementary data to a lexical production task, without however a major impact on the evaluation of the systems.

Lexical and Semantic Features for Cross-lingual Text Reuse Classification: an Experiment in English and Latin Paraphrases

Maria Moritz and David Steding

Analyzing historical languages, such as Ancient Greek and Latin, is challenging. Such languages are often under-resourced and lack primary material for certain time periods. This prevents applying advanced natural-language processing (NLP) techniques and requires resorting to basic NLP not relying on machine learning. An important analysis is the discovery and classification of paraphrastic text reuse in historical languages. This reuse is often paraphrastic and challenges basic NLP techniques. Our goal is to improve the applicability of advanced NLP techniques on historical text reuse. We present an experiment of cross-applying classifiers—that we trained for paraphrase recognition on modern English text corpora—on historical texts. We analyze the impact of four different lexical and semantic features, on the resulting reuse-detection accuracy. We find out that—against initial conjecture—word embedding can help to drastically improve accuracy if lexical features (such as the overlap of similar words) fail.

Investigating the Influence of Bilingual MWU on Trainee Translation Quality

YU Yuan and Serge Sharoff

We applied a method for automatic extraction of bilingual multiword units (BMWUs) from a parallel corpus in order to investigate their contribution to translation quality in terms of adequacy and fluency. Our statistical analysis is based on generalized additive modelling. It has been shown that normalised BMWU ratios can be useful for estimating human translation quality. The normalized alignment ratios for BMWUs longer than two words have the greatest impact on measuring translation quality. It is also found that the alignment ratio for longer BMWUs is statistically closer to adequacy than to fluency.

Evaluation of Dictionary Creating Methods for Finno-Ugric Minority Languages

Zsanett Ferenczi, Iván Mittelholcz, Eszter Simon and Tamás Váradi

In this paper, we present the evaluation of several bilingual dictionary building methods applied to {Komi-Permyak, Komi-Zyrian, Hill Mari, Meadow Mari, Northern Saami, Udmurt}-{English, Finnish, Hungarian, Russian} language pairs. Since these Finno-Ugric minority languages are under-resourced and standard dictionary building methods require a large amount of pre-processed data, we had to find alternative methods. In a thorough evaluation, we compare the results for each method, which proved our expectations that the precision of standard lexicon building methods is quite low for under-resourced

languages. However, utilizing Wikipedia title pairs extracted via inter-language links and Wiktionary-based methods provided useful results. The newly created word pairs enriched with several linguistic information are to be deployed on the web in the framework of Wiktionary. With our dictionaries, the number of Wiktionary entries in the above mentioned Finno-Ugric minority languages can be multiplied.

Dysarthric speech evaluation: automatic and perceptual approaches

Imed Laaridh, Christine Meunier and Corinne Fredouille

Perceptual evaluation is still the most common method in clinical practice for the diagnosis and monitoring of the condition progression of people suffering from dysarthria (or speech disorders more generally). Such evaluations are frequently described as non-trivial, subjective and highly time-consuming (depending on the evaluation level). Clinicians have, therefore, expressed their need for new objective evaluation tools more adapted to longitudinal studies or rehabilitation context. We proposed earlier an automatic approach for the anomaly detection at the phone level for dysarthric speech. The system behavior was studied and validated on different corpora and speech styles. Nonetheless, the lack of annotated French dysarthric speech corpora has limited our capacity to analyze some aspects of its behavior, and notably,its severity (more anomalies detected automatically compared with human expert). To overcome this limitation, we proposed an original perceptual evaluation protocol applied to a limited set of decisions made by the automatic system, related to the presence of anomalies. This evaluation was carried out by a jury of 29 non-naive individuals. In addition to interesting information related to the system behavior, the evaluation protocol highlighted the difficulty for a human, even expert, to apprehend and detect deviations at the word level in dysarthric speech.

Towards an Automatic Assessment of Crowdsourced Data for NLU

Patricia Braunger, Wolfgang Maier, Jan Wessling and Maria Schmidt

Recent development of spoken dialog systems has moved away from a command-style input and aims at allowing a natural input style. Obtaining suitable data for training and testing such systems is a significant challenge. We investigate with which methods data elicited via crowdsourcing can be assessed with respect to its naturalness and usefulness. Since the criteria with which to assess usefulness depend on the application purpose of crowdsourced data we investigate various facets such as noisy data, naturalness and building natural language understanding (NLU) models. Our results show that valid data can be automatically identified with the help of a word based language model. A comparison of crowdsourced data and system usage data on lexical, syntactic and pragmatic level reveals detailed information on the differences between both data sets. However, we show that using crowdsourced data for training NLU services achieves similar results as system usage data.

Visual Choice of Plausible Alternatives: An Evaluation of Image-based Commonsense Causal Reasoning

Jinyoung Yeo, Gyeongbok Lee, Gengyu Wang, Seungtaek Choi, Hyunsouk Cho, Reinald Kim Amplayo and Seungwon Hwang

This paper proposes the task of Visual COPA (VCOPA). Given a premise image and two alternative images, the task is to identify the more plausible alternative with their commonsense causal context. The VCOPA task is designed as its desirable machine system needs a more detailed understanding of the image, commonsense knowledge, and complex causal reasoning than state-of-the-art AI techniques. For that, we generate an evaluation dataset containing 380 VCOPA questions and over 1K images with various topics, which is amenable to automatic evaluation, and present the performance of baseline reasoning approaches as initial benchmarks for future systems.

Is it worth it? Budget-related evaluation metrics for model selection

Filip Klubička, Giancarlo D. Salton and John D. Kelleher

Creating a linguistic resource is often done by using a machine learning model that filters the content that goes through to a human annotator, before going into the final resource. However, budgets are often limited, and the amount of available data exceeds the amount of affordable annotation. In order to optimize the benefit from the invested human work, we argue that deciding on which model one should employ depends not only on generalized evaluation metrics such as F-score, but also on the gain metric. Because the model with the highest F-score may not necessarily have the best sequencing of predicted classes, this may lead to wasting funds on annotating false positives, yielding zero improvement of the linguistic resource. We exemplify our point with a case study, using real data from a task of building a verb-noun idiom dictionary. We show that, given the choice of three systems with varying F-scores, the system with the highest F-score does not yield the highest profits. In other words, in our case the cost-benefit trade off is more favorable for a system with a lower F-score.

Automated Evaluation of Out-of-Context Errors

Patrick Huber, Jan Niehues and Alex Waibel

We present a new approach to evaluate computational models for the task of text understanding by the means of out-of-context error detection. Through the novel design of our automated modification process, existing large-scale data sources can be adopted for a vast number of text understanding tasks. The data is thereby altered on a semantic level, allowing models to be tested against a challenging set of modified text passages that require to comprise a broader narrative discourse. Our newly introduced task targets actual real-world problems of transcription and translation systems by inserting authentic out-of-context errors. The automated modification process is applied to the 2016 TEDTalk corpus. Entirely automating the process allows the adoption of complete datasets at low cost, facilitating supervised learning procedures and deeper networks to be trained and tested. To evaluate the quality of the modification algorithm a language model and a supervised binary classification model are trained and tested on the altered dataset. A human baseline evaluation is examined to compare the results with human performance. The outcome of the evaluation task indicates the difficulty to detect semantic errors for machine-learning algorithms and humans, showing that the errors cannot be identified when limited to a single sentence.

Matics Software Suite: New Tools for Evaluation and Data Exploration

Olivier Galibert, Guillaume Bernard, Agnes Delaborde, Sabrina Lecadre and Juliette Kahn

Matics is a free and open-source software suite for exploring annotated data and evaluation results. It proposes a dataframe data model allowing the intuitive exploration of data characteristics and evaluation results and provides support for graphing the values and running appropriate statistical tests. The tools already run on several Natural Language Processing tasks and standard annotation formats, and are under on-going development.

MGAD: Multilingual Generation of Analogy Datasets

Mostafa Abdou, Artur Kulmizev and Vinit Ravishankar

We present a novel, minimally supervised method of generating word embedding evaluation datasets for a large number of languages. Our approach utilizes existing dependency treebanks and parsers in order to create language-specific syntactic analogy datasets that do not rely on translation or human annotation. As part of our work, we offer syntactic analogy datasets for three previously unexplored languages: Arabic, Hindi, and Russian. We further present an evaluation of three popular word embedding algorithms (Word2Vec,GloVe, LexVec) against these datasets and explore how the performance of each word embedding algorithm varies between several syntactic categories.

Session P23 - Information Extraction, Information Retrieval, Text Analytics (2)

10th May 2018, 09:45

Chair person: **Pierre Zweigenbaum** Poster Session

MIsA: Multilingual "IsA" Extraction from Corpora

Stefano Faralli, Els Lefever and Simone Paolo Ponzetto

In this paper we present Multilingual IsA (MIsA), which is a collection of hypernymy relations in five languages (i.e., English, Spanish, French, Italian and Dutch) extracted from the corresponding full Wikipedia corpus. For each language, we first established a set of existing (viz. found in literature) or newly defined lexico-syntactic patterns. Similarly to WebIsADb, the resulting resource contains hypernymy relations represented as "tuples", as well as additional information such as provenance, context of the extraction, etc. To measure the precision of the patterns, we performed a manual assessment of the quality of the extracted relations and an error analysis. In addition, we release the software developed for the extraction of the hypernym tuples.

Biomedical term normalization of EHRs with UMLS

Naiara Perez, Montse Cuadros and German Rigau

This paper presents a novel prototype for biomedical term normalization of electronic health record excerpts with the Unified Medical Language System (UMLS) Metathesaurus, a large, multi-lingual compendium of biomedical and health-related terminologies. Despite the prototype being multilingual and cross-lingual by design, we first focus on processing clinical text in Spanish because there is no existing tool for this language and for this specific purpose. The tool is based on Apache Lucene to index the Metathesaurus and generate mapping candidates from input text. It uses the IXA pipeline for basic language processing and resolves lexical ambiguities with the UKB toolkit. It has been evaluated by measuring its agreement with MetaMap –a mature software to discover UMLS concepts in English texts– in two English-Spanish parallel corpora. In addition, we present a web-based interface for the tool.

Revisiting the Task of Scoring Open IE Relations

William Lechelle and Phillippe Langlais

Knowledge Base Completion infers missing facts from existing ones in knowledge bases. As recent Open Information Extraction systems allow us to extract ever larger (yet incomplete) open-domain Knowledge Bases from text, we seek to probabilistically extend the limited coverage we get from existing facts, to arbitrary queries about plausible information. We propose a simple baseline, based on language modeling and trained with off-the-shelf programs, which gives competitive results in the previously defined protocol for this task, and provides an independent source of signal to judge arbitrary fact plausibility. We reexamine this protocol, measure the (large) impact of the negative example generation procedure, which we find to run contrary to the belief put forward in previous work. We conduct a small manual evaluation, giving insights into the rudimentary automatic evaluation protocol, and analyse the shortcomings of our model.

A supervised approach to taxonomy extraction using word embeddings

Rajdeep Sarkar, John Philip McCrae and Paul Buitelaar

Large collections of texts are commonly generated by large organizations and making sense of these collections of texts is a significant challenge. One method for handling this is to organize the concepts into a hierarchical structure such that similar concepts can be discovered and easily browsed. This approach was the subject of a recent evaluation campaign, TExEval, however the results of this task showed that none of the systems consistently outperformed a relatively simple baseline.In order to solve this issue, we propose a new method that uses supervised learning to combine multiple features with a support vector machine classifier including the baseline features. We show that this outperforms the baseline and thus provides a stronger method for identifying taxonomic relations than previous methods

A Chinese Dataset with Negative Full Forms for General Abbreviation Prediction

Yi Zhang and Sun Xu

Abbreviation is a common phenomenon across languages, especially in Chinese. In most cases, if an expression can be abbreviated, its abbreviation is used more often than its fully expanded forms, since people tend to convey information in a most concise way. For various language processing tasks, abbreviation is an obstacle to improving the performance, as the textual form of an abbreviation does not express useful information, unless it's expanded to the full form. Abbreviation prediction means associating the fully expanded forms with their abbreviations. However, due to the deficiency in the abbreviation corpora, such a task is limited in current studies, especially considering general abbreviation prediction should also include those full form expressions that do not have general abbreviations, namely the negative full forms (NFFs). Corpora incorporating negative full forms for general abbreviation prediction are few in number. In order to promote the research in this area, we build a dataset for general Chinese abbreviation prediction, which needs a few preprocessing steps, and evaluate several different models on the built dataset. The dataset is available at https://github.com/lancopku/Chinese-abbreviation-dataset.

Korean TimeBank Including Relative Temporal Information

Chae-Gyun Lim, Young-Seob Jeong and Ho-Jin Choi

Since most documents have temporal information that can be a basis for understanding the context, the importance of temporal information extraction researches is steadily growing. Although various attempts have been made to extract temporal information from researchers internationally, it is difficult to apply them to other languages because they are usually targeted at specific languages such as English. Several annotation languages and datasets had been proposed for the studies of temporal information extraction on Korean documents, however, the representation of relative temporal information is not enough to maintain it explicitly. In this paper, we propose a concept of relative temporal information and supplement a Korean annotation language to represent new relative expressions, and extend an annotated dataset, Korean TimeBank, through the revised language. We expect that it is possible to utilize potential features from the Korean corpus by clearly annotating relative temporal relationships and to use well-refined Korean TimeBank in future studies.

Mining Biomedical Publications With The LAPPS Grid

Nancy Ide, Keith Suderman and Jin-Dong Kim

It is widely recognized that the ability to exploit Natural Language Processing (NLP) text mining strategies has the potential to increase productivity and innovation in the sciences by orders of magnitude, by enabling scientists to pull information from research articles in scientific disciplines such as genomics and biomedicine. The Language Applications (LAPPS) Grid is an infrastructure for rapid development of natural language processing applications (NLP) that provides an ideal platform to support mining scientific literature. Its Galaxy interface and the interoperability among tools together provide an intuitive and easy-to-use platform, and users can experiment with and

exploit NLP tools and resources without the need to determine which are suited to a particular task, and without the need for significant computer expertise. The LAPPS Grid has collaborated with the developers of PubAnnotation to integrate the services and resources provided by each in order to greatly enhance the user's ability to annotate scientific publications and share the results. This poster/demo shows how the LAPPS Grid can facilitate mining scientific publications, including identification and extraction of relevant entities, relations, and events; iterative manual correction and evaluation of automatically-produced annotations, and customization of supporting resources to accommodate specific domains.

An Initial Test Collection for Ranked Retrieval of SMS Conversations

Rashmi Sankepally and Douglas W. Oard

This paper describes a test collection for evaluating systems that search English SMS (Short Message Service) conversations. The collection is built from about 120,000 text messages. Topic development involved identifying typical types of information needs, then generating topics of each type for which relevant content might be found in the collection. Relevance judgments were then made for groups of messages that were most highly ranked by one or more of several ranked retrieval systems. The resulting TREC style test collection can be used to compare some alternative retrieval system designs.

FrNewsLink : a corpus linking TV Broadcast News Segments and Press Articles

Nathalie Camelin, Géraldine Damnati, Abdessalam Bouchekif, Anais Landeau, Delphine Charlet and Yannick Esteve

In this article, we describe FrNewsLink, a corpus allowing to address several applicative tasks that we make publicly available. It gathers several resources from TV Broadcast News (TVBN) shows and press articles such as automatic transcription of TVBN shows, text extracted from on-line press articles, manual annotations for topic segmentation of TVBN shows and linking information between topic segments and press articles. The FrNewsLink corpus is based on 112 (TVBN) shows recorded during two periods in 2014 and 2015. Concomitantly, a set of 24,7k press articles has been gathered. Beyond topic segmentation, this corpus allows to study semantic similarity and multimedia News linking.

PyRATA, Python Rule-based feAture sTructure Analysis

Nicolas Hernandez and Amir Hazem

In this paper, we present a new Python 3 module named PyRATA, which stands for "Python Rule-based feAture sTructure Analysis". The module is released under the Apache V2 license. It aims at supporting rules-based analysis on structured data. PyRATA offers a language expressiveness which covers the functionalities of all the concurrent modules and more. Designed to be intuitive, the pattern syntax and the engine API follow existing standard definitions; Respectively Perl regular expression syntax and Python re module API. Using a simple native Python data structure (i.e. sequence of feature sets) allows it to deal with various kinds of data (textual or not) at various levels, such as a list of words, a list of sentences, a list of posts of a forum thread, a list of events of a calendar... This specificity makes it free from any (linguistic) process.

Session P24 - Multimodality

10th May 2018, 09:45

Chair person: **Martin Braschler** Poster Session

Towards Processing of the Oral History Interviews and Related Printed Documents

Zbynek Zajic, Lucie Skorkovska, Petr Neduchal, Pavel Ircing, Josef V. Psutka, Marek Hruz, Ales Prazak, Daniel Soutner, Jan Švec, Lukas Bures and Ludek Muller

In this paper, we describe the initial stages of our project, the goal of which is to create an integrated archive of the recordings, scanned documents, and photographs that would be accessible online and would provide multifaceted search capabilities (spoken content, biographical information, relevant time period, etc.). The recordings contain retrospective interviews with the witnesses of the totalitarian regimes in Czechoslovakia, where the vocabulary used in such interviews consists of many archaic words and named entities that are now quite rare in everyday speech. The scanned documents consist of text materials and photographs mainly from the home archives of the interviewees or the archive of the State Security. These documents are usually typewritten or even handwritten and have really bad optical quality. In order to build an integrated archive, we will employ mainly methods of automatic speech recognition (ASR), automatic indexing and search in recognized recordings and, to a certain extent, also the optical character recognition (OCR). Other natural language processing techniques like topic detection are also planned to be used in the later stages of the project. This paper focuses on the processing of the speech data using ASR and the

scanned typewritten documents with OCR and describes the initial experiments.

Multi Modal Distance - An Approach to Stemma Generation With Weighting

Armin Hoenen

Stemma generation can be understood as a task where an original manuscript M gets copied and copies – due to the manual mode of copying – vary from each other and from M . Copies $M_1 , .. , M_k$ which survive historical loss serve as input to a mapping process estimating a directed acyclic graph (tree) which is the most likely representation of their copy history. One can first tokenize and align the texts of $M_1, .., M_k$ and then produce a pairwise distance matrix between them. From this, one can finally derive a tree with various methods, for instance Neighbor-Joining (NJ) (Saitou and Nei, 1987). For computing those matrices, previous research has applied unweighted approaches to token similarity (implicitly interpreting each token pair as a binary observation: identical or different), see Mooney et al. (2003). The effects of weighting have then been investigated and Spencer et al. (2004b) found them to be small in their (not necessarily all) scenario(s). The present approach goes beyond the token level and instead of a binary comparison uses a distance model on the basis of psycholinguistically gained distance matrices of letters in three modalities: vision, audition and motorics. Results indicate that this type of weighting have positive effects on stemma generation.

A Corpus of Natural Multimodal Spatial Scene Descriptions

Ting Han and David Schlangen

We present a corpus of multimodal spatial descriptions, a common scenario in route giving tasks. Participants provided natural spatial scene descriptions with speech and iconic/abstract deictic hand gestures. The scenes were composed of simple geometric objects. While the language denotes object shape and visual properties (e.g., colour), the abstract deictic gestures "placed" objects in gesture space to denote spatial relations of objects. Only together with speech do these gestures receive defined meanings. Hence, the presented corpus goes beyond previous work on gestures in multimodal interfaces that either focusses on gestures with predefined meanings (multimodal commands) or provides hand motion data without accompanying speech. At the same time, the setting is more constrained than full human/human interaction, making the resulting data more amenable to computational analysis and more directly useable for learning natural computer interfaces. Our preliminary analysis results show that co-verbal deictic gestures in the corpus reflect spatial configurations of objects, and there are variations of gesture space and verbal descriptions. The provided verbal descriptions and hand motion data will enable modelling the interpretations of natural multimodal descriptions with machine learning methods, as well as other tasks such as generating natural multimodal spatial descriptions.

The Effects of Unimodal Representation Choices on Multimodal Learning

Fernando T. Ito, Helena de Medeiros Caseli and Jander Moreira

Multimodal representations are distributed vectors that map multiple modes of information to a single mathematical space, where distances between instances delineate their similarity. In most cases, using a single unimodal representation technique is sufficient for each mode in the creation of multimodal spaces. In this paper, we investigate how different unimodal representations can be combined, and argue that the way they are combined can affect the performance, representation accuracy and classification metrics of other multimodal methods. In the experiments present in this paper, we used a dataset composed of images and text descriptions of products that have been extracted from an e-commerce site in Brazil. From this dataset, we tested our hypothesis in common classification problems to evaluate how multimodal representations can differ according to their component unimodal representation methods. For this domain, we selected eight methods of unimodal representation: LSI, LDA, Word2Vec, GloVe for text; SIFT, SURF, ORB and VGG19 for images. Multimodal representations were built by a multimodal deep autoencoder and a bidirectional deep neural network.

An Evaluation Framework for Multimodal Interaction

Nikhil Krishnaswamy and James Pustejovsky

In this paper we present a framework for evaluating interactions between a human user and an embodied virtual agent that communicates using natural language, gesture, and by executing actions in a shared context created through a visual simulation interface. These interactions take place in real time and demonstrate collaboration between a human and a computer on object interaction and manipulation. Our framework leverages the semantics of language and gesture to assess the level of mutual understanding during the interaction and the ease with which the two agents communicate. We present initial results from trials involving construction tasks in a blocks world scenario and discuss extensions of the evaluation framework to more naturalistic and robust interactions.

The WAW Corpus: The First Corpus of Interpreted Speeches and their Translations for English and Arabic

Ahmed Abdelali, Irina Temnikova, Samy Hedaya and Stephan Vogel

This article presents the WAW Corpus, an interpreting corpus for English/Arabic, which can be used for teaching interpreters, studying the characteristics of interpreters' work, as well as to train machine translation systems. The corpus contains recordings of lectures and speeches from international conferences, their interpretations, the transcripts of the original speeches and of their interpretations, as well as human translations of both kinds of transcripts into the opposite language of the language pair. The article presents the corpus curation, statistics, assessment, as well as a case study of the corpus use.

Polish Corpus of Annotated Descriptions of Images

Alina Wróblewska

The paper presents a new dataset of image descriptions in Polish. The descriptions are morphosyntactically analysed and the pairs of these descriptions are annotated in terms of semantic relatedness and entailment. All annotations are provided by human annotators with strong linguistic background. The dataset can be used for evaluation of various systems integrating language and vision. It is applicable for evaluation of systems designed to image generation based on provided descriptions (text-to-image generation) or to caption generation based on images (image-to-text generation). Furthermore, as selected images are split into thematic groups, the dataset is also useful for validating image classification approaches.

Action Verb Corpus

Stephanie Gross, Matthias Hirschmanner, Brigitte Krenn, Friedrich Neubarth and Michael Zillich

The Action Verb Corpus comprises multimodal data of 12 humans conducting in total 390 simple actions (take, put, and push). Recorded are audio, video and motion data while participants perform an action and describe what they do. The dataset is annotated with the following information: orthographic transcriptions of utterances, part-of-speech tags, lemmata, information which object is currently moved, information whether a hand touches an object, information whether an object touches the ground/table. Transcription, and information whether an object is in contact with a hand and which object moves where to were manually annotated, the rest was automatically annotated and manually corrected. In addition to the dataset, we present an algorithm for the challenging task of segmenting the stream of words into utterances, segmenting the visual input into a series of actions, and then aligning visual action information and speech. This kind of modality rich data is particularly important for crossmodal and cross-situational word-object and word-action learning in human-robot interactions, and is comparable to parent-toddler communication in early stages of child language acquisition.

EMO&LY (EMOtion and AnomaLY) : A new corpus for anomaly detection in an audiovisual stream with emotional context.

Cédric Fayet, Arnaud Delhay, Damien Lolive and Pierre-françois Marteau

This paper presents a new corpus, called EMOLY (EMOtion and AnomaLY), composed of speech and facial video records of subjects that contains controlled anomalies. As far as we know, to study the problem of anomaly detection in discourse by using machine learning classification techniques, no such corpus exists or is available to the community. In EMOLY, each subject is recorded three times in a recording studio, by filming his/her face and recording his/her voice with a HiFi microphone. Anomalies in discourse are induced or acted. At this time, about 8,65 hours of usable audiovisual recording on which we have tested classical classification techniques (GMM or One Class-SVM plus threshold classifier) are available. Results confirm the usability of the anomaly induction mechanism to produce anomalies in discourse and also the usability of the corpus to improve detection techniques.

Development of an Annotated Multimodal Dataset for the Investigation of Classification and Summarisation of Presentations using High-Level Paralinguistic Features

Keith Curtis, Nick Campbell and Gareth Jones

Expanding online archives of presentation recordings provide potentially valuable resources for learning and research. However, the huge volume of data that is becoming available means that users have difficulty locating material which will be of most value to them. Conventional summarisation methods making use of text-based features derived from transcripts of spoken material can provide mechanisms to rapidly locate topically interesting material by reducing the amount of material that must be auditioned. However, these text-based methods take no account of the multimodal high-level paralinguistic features which form part of an audio-visual presentation, and can provide valuable indicators of the most interesting material within a presentation. We describe the development of a multimodal video dataset, recorded at an international conference, designed to support the

exploration of automatic extraction of paralinguistic features and summarisation based on these features. The dataset is comprised of parallel recordings of the presenter and the audience for 31 conference presentations. We describe the process of performing manual annotation of high-level paralinguistic features for speaker ratings, audience engagement, speaker emphasis, and audience comprehension of these recordings. Used in combination these annotations enable research into the automatic classification of high-level paralinguistic features and their use in video summarisation.

Session P25 - Parsing, Syntax, Treebank (1)

10th May 2018, 09:45

Chair person: **Simonetta Montemagni** Poster Session

BKTreebank: Building a Vietnamese Dependency Treebank

Kiem-Hieu Nguyen

Dependency treebank is an important resource in any language. In this paper, we present our work on building BKTreebank, a dependency treebank for Vietnamese. Important points on designing POS tagset, dependency relations, and annotation guidelines are discussed. We describe experiments on POS tagging and dependency parsing on the treebank. Experimental results show that the treebank is a useful resource for Vietnamese language processing.

GeCoTagger: Annotation of German Verb Complements with Conditional Random Fields

Roman Schneider and Monica Fürbacher

Complement phrases are essential for constructing well-formed sentences in German. Identifying verb complements and categorizing complement classes is challenging even for linguists who are specialized in the field of verb valency. Against this background, we introduce an ML-based algorithm which is able to identify and classify complement phrases of any German verb in any written sentence context. We use a large training set consisting of example sentences from a valency dictionary, enriched with POS tagging, and the ML-based technique of Conditional Random Fields (CRF) to generate the classification models.

AET: Web-based Adjective Exploration Tool for German

Tatiana Bladier, Esther Seyffarth, Oliver Hellwig and Wiebke Petersen

We present a new web-based corpus query tool, the Adjective Exploration Tool (AET), which enables research on the modificational behavior of German adjectives and adverbs. The tool can also be transferred to other languages and modification phenomena. The underlying database is derived from a corpus of German print media texts, which we annnotated with dependency parses and several morphological, lexical, and statistical properties of the tokens. We extracted pairs of adjectives and adverbs (modifiers) as well as the tokens modified by them (modifiees) from the corpus and stored them in a way that makes the modifier-modifiee pairs easily searchable. With AET, linguists from different research areas can access corpus samples using an intuitive query language and user-friendly web interface. AET has been developed as a part of a collaborative research project that focuses on the compositional interaction of attributive adjectives with nouns and the interplay of events and adverbial modifiers. The tool is easy to extend and update and is free to use online without registration: http://aet.phil.hhu.de

ZAP: An Open-Source Multilingual Annotation Projection Framework

Alan Akbik and Roland Vollgraf

Previous work leveraged annotation projection as a convenient method to automatically generate linguistic resources such as treebanks or propbanks for new languages. This approach automatically transfers linguistic annotation from a resource-rich source language (SL) to translations in a target language (TL). However, to the best of our knowledge, no publicly available framework for this approach currently exists, limiting researchers' ability to reproduce and compare experiments. In this paper, we present ZAP, the first open-source framework for annotation projection in parallel corpora. Our framework is Java-based and includes methods for preprocessing corpora, computing word-alignments between sentence pairs, transferring different layers of linguistic annotation, and visualization. The framework was designed for ease-of-use with lightweight APIs. We give an overview of ZAP and illustrate its usage.

Palmyra: A Platform Independent Dependency Annotation Tool for Morphologically Rich Languages

Talha Javed, Nizar Habash and Dima Taji

We present Palmyra, a platform independent graphical dependency tree visualization and editing software. Palmyra has been specifically designed to support the complexities of syntactic annotation of morphologically rich languages, especially regarding easy change of morphological tokenization through edits, additions, deletions, splits and merges of words. Palmyra uses an intuitive drag-and-drop interface for editing tree structures, and provides pop-up boxes and keyboard shortcuts for part-of-speech and link label tagging.

A Web-based System for Crowd-in-the-Loop Dependency Treebanking

Stephen Tratz and Nhien Phan

Treebanks exist for many different languages, but they are often quite limited in terms of size, genre, and topic coverage. It is difficult to expand these treebanks or to develop new ones in part because manual annotation is time-consuming and expensive. Human-in-the-loop methods that leverage machine learning algorithms during the annotation process are one set of techniques that could be employed to accelerate annotation of large numbers of sentences. Additionally, crowdsourcing could be used to hire a large number of annotators at relatively low cost. Currently, there are few treebanking tools available that support either human-in-the-loop methods or crowdsourcing. To address this, we introduce CrowdTree , a web-based interactive tool for editing dependency trees. In addition to the visual frontend, the system has a Java servlet that can train a parsing model during the annotation process. This parsing model can then be applied to sentences as they are requested by annotators so that, instead of annotating sentences from scratch, annotators need only to edit the model's predictions, potentially resulting in significant time savings. Multiple annotators can work simultaneously, and the system is even designed to be compatible with Mechanical Turk. Thus, CrowdTree supports not simply human-in-the-loop treebanking, but crowd-in-the-loop treebanking.

Building Universal Dependency Treebanks in Korean

Jayeol Chun, Na-Rae Han, Jena D. Hwang and Jinho D. Choi

This paper presents three treebanks in Korean that consist of dependency trees derived from existing treebanks, the Google UD Treebank, the Penn Korean Treebank, and the Kaist Treebank, and pseudo-annotated by the latest guidelines from the Universal Dependencies (UD) project. The Korean portion of the Google UD Treebank is re-tokenized to match the morpheme-level annotation suggested by the other corpora, and systematically assessed for errors. Phrase structure trees in the Penn Korean Treebank and the Kaist Treebank are automatically converted into dependency trees using head-finding rules and linguistic heuristics. Additionally, part-of-speech tags in all treebanks are converted into the UD tagset. A total of 38K+ dependency trees are generated that comprise a coherent set of dependency relations for over a half million tokens. To the best of our knowledge, this is the first time that these Korean corpora are analyzed together and transformed into dependency trees following the latest UD guidelines, version 2.

Moving TIGER beyond Sentence-Level

Agnieszka Falenska, Kerstin Eckart and Jonas Kuhn

We present TIGER 2.2-doc – a new set of annotations for the German TIGER corpus. The set moves the corpus to a document level. It includes a full mapping of sentences to documents, as well as additional sentence-level and document-level annotations. The sentence-level annotations refer to the role of a sentence in the document. They introduce structure to the TIGER documents by separating headers and meta-level information from article content. Document-level annotations recover information which has been neglected in the intermediate releases of the TIGER corpus, such as document categories and publication dates of the articles. Additionally, we introduce new document-level annotations: authors and their gender. We describe the process of corpus annotation, show statistics of the obtained data and present baseline experiments for lemmatization, part-of-speech and morphological tagging, and dependency parsing. Finally, we present two example use cases: sentence boundary detection and authorship attribution. These use cases take the data from TIGER into account and illustrate the usefulness of the new annotation layers from TIGER 2.2-doc.

Spanish HPSG Treebank based on the AnCora Corpus

Luis Chiruzzo and Dina Wonsever

This paper describes a corpus of HPSG annotated trees for Spanish that contains morphosyntactic information, annotations for semantic roles, clitic pronouns and relative clauses. The corpus is based on the Spanish AnCora corpus, which contains trees for 17,000 sentences comprising half a million words, and it has CFG style annotations. The corpus is stored in two different formats: An XML dialect that is the direct serialization of the typed feature structure trees, and an HTML format that is suitable for visualizing the trees in a browser.

Universal Dependencies for Amharic

Binyam Ephrem Seyoum, Yusuke Miyao and Baye Yimam Mekonnen

In this paper, we describe the process of creating an Amharic Dependency Treebank, which is the first attempt to introduce Universal Dependencies (UD) into Amharic. Amharic is a morphologically-rich and less-resourced language within the Semitic language family. In Amharic, an orthographic word may be bundled with information other than morphology. There are some clitics attached to major lexical categories with grammatical functions. We first explain the segmentation of clitics, which is problematic to retrieve from the orthographic word due to morpheme co-occurrence restriction, assimilation and ambiguity of the clitics. Then, we describe the annotation processes for POS

tagging, morphological information and dependency relations. Based on this, we have created a Treebank of 1,096 sentences.

A Parser for LTAG and Frame Semantics

David Arps and Simon Petitjean

Since the idea of combining Lexicalized Tree Adjoining Grammars (LTAG) and frame semantics was proposed (Kallmeyer and Osswald, 2013), a set of resources of this type has been created. These grammars are composed of pairs of elementary trees and frames, where syntactic and semantic arguments are linked using unification variables. This allows to build semantic representations when parsing, by composing the frames according to the combination of the elementary trees. However, the lack of a parser using such grammars makes it complicated to check whether these resources are correct implementations of the theory or not. The development of larger resources, that is to say large-coverage grammars, is also conditioned by the existence of such a parser. In this paper, we present our solution to this problem, namely an extension of the TuLiPA parser with frame semantics. We also present the frameworks used to build the resources used by the parser: the theoretical framework, composed of LTAG and frame semantics, and the software framework, XMG2.

Multilingual Dependency Parsing for Low-Resource Languages: Case Studies on North Saami and Komi-Zyrian

KyungTae Lim, Niko Partanen and Thierry Poibeau

The paper presents a method for parsing low-resource languages with very small training corpora using multilingual word embeddings and annotated corpora of larger languages. The study demonstrates that specific language combinations enable improved dependency parsing when compared to previous work, allowing for wider reuse of pre-existing resources when parsing low-resource languages. The study also explores the question of whether contemporary contact languages or genetically related languages would be the most fruitful starting point for multilingual parsing scenarios.

Session O17 - Evaluation Methodologies

10th May 2018, 11:45

Chair person: **Maciej Piasecki** Oral Session

Evaluating the WordsEye Text-to-Scene System: Imaginative and Realistic Sentences

Morgan Ulinski, Bob Coyne and Julia Hirschberg

We describe our evaluation of the WordsEye text-to-scene generation system. We address the problem of evaluating the output of such a system vs. simple search methods to find a picture to illustrate a sentence. To do this, we constructed two sets of test sentences: a set of crowdsourced imaginative sentences and a set of realistic sentences extracted from the PASCAL image caption corpus (Rashtchian et al., 2010). For each sentence, we compared sample pictures found using Google Image Search to those produced by WordsEye. We then crowdsourced judgments as to which picture best illustrated each sentence. For imaginative sentences, pictures produced by WordsEye were preferred, but for realistic sentences, Google Image Search results were preferred. We also used crowdsourcing to obtain a rating for how well each picture illustrated the sentence, from 1 (completely correct) to 5 (completely incorrect). WordsEye pictures had an average rating of 2.58 on imaginative sentences and 2.54 on realistic sentences; Google images had an average rating of 3.82 on imaginative sentences and 1.87 on realistic sentences. We also discuss the sources of errors in theWordsEye system.

Computer-assisted Speaker Diarization: How to Evaluate Human Corrections

Pierre-Alexandre Broux, David Doukhan, Simon Petitrenaud, Sylvain Meignier and Jean Carrive

In this paper, we present a framework to evaluate the human correction of a speaker diarization. We propose four elementary actions to correct the diarization and an automaton to simulate the correction sequence. A metric is described to evaluate the correction cost. The framework is evaluated using French broadcast news drawn from the REPERE corpus.

Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment

Masatoshi Tsuchiya

The quality of training data is one of the crucial problems when a learning-centered approach is employed. This paper proposes a new method to investigate the quality of a large corpus designed for the recognizing textual entailment (RTE) task. The proposed method, which is inspired by a statistical hypothesis test, consists of two phases: the first phase is to introduce the predictability of textual entailment labels as a null hypothesis which is extremely unacceptable if a target corpus has no hidden bias, and the second phase is to test the null hypothesis using a Naive Bayes model. The experimental result of the Stanford Natural Language Inference (SNLI) corpus does not reject the null hypothesis. Therefore, it indicates that the SNLI corpus has a hidden bias which allows prediction of textual entailment labels from hypothesis sentences even if no context information is given by a premise sentence. This paper also presents the performance impact of NN models for RTE caused by this hidden bias.

Evaluation of Croatian Word Embeddings

Lukas Svoboda and Slobodan Beliga

Croatian is poorly resourced and highly inflected language from Slavic language family. Nowadays, research is focusing mostly on English. We created a new word analogy dataset based on the original English Word2vec word analogy dataset and added some of the specific linguistic aspects from Croatian language. Next, we created Croatian WordSim353 and RG65 datasets for a basic evaluation of word similarities. We compared created datasets on two popular word representation models, based on Word2Vec tool and fastText tool. Models has been trained on 1.37B tokens training data corpus and tested on a new robust Croatian word analogy dataset. Results show that models are able to create meaningful word representation. This research has shown that free word order and the higher morphological complexity of Croatian language influences the quality of resulting word embeddings.

Session O18 - Semantics

10th May 2018, 11:45

Chair person: **James Pustejovsky** Oral Session

C-HTS: A Concept-based Hierarchical Text Segmentation approach

Mostafa Bayomi and Seamus Lawless

Hierarchical Text Segmentation is the task of building a hierarchical structure out of text to reflect its sub-topic hierarchy. Current text segmentation approaches are based upon using lexical and/or syntactic similarity to identify the coherent segments of text. However, the relationship between segments may be semantic, rather than lexical or syntactic. In this paper we propose C-HTS, a Concept-based Hierarchical Text Segmentation approach that uses the semantic relatedness between text constituents. In this approach, we use the explicit semantic representation of text, a method that replaces keyword-based text representation with concept-based features, automatically extracted from massive human knowledge repositories such as Wikipedia. C-HTS represents the meaning of a piece of text as a weighted vector of knowledge concepts, in order to reason about text. We evaluate the performance of C-HTS on two publicly available datasets. The results show that C-HTS compares favourably with previous state-of-the-art approaches. As Wikipedia is continuously growing, we measured the impact of its growth on segmentation performance. We used three different snapshots of Wikipedia from different years in order to achieve this. The experimental results show that an increase in the size of the knowledge base leads, on average, to greater improvements in hierarchical text segmentation.

Semantic Supersenses for English Possessives

Austin Blodgett and Nathan Schneider

We adapt an approach to annotating the semantics of adpositions to also include English possessives, showing that the supersense inventory of Schneider et al. (2017) works for the genitive 's clitic and possessive pronouns as well as prepositional of. By comprehensively annotating such possessives in an English corpus of web reviews, we demonstrate that the existing supersense categories are readily applicable to possessives. Our corpus will facilitate empirical study of the semantics of the genitive alternation and the development of semantic disambiguation systems.

A Corpus of Metaphor Novelty Scores for Syntactically-Related Word Pairs

Natalie Parde and Rodney Nielsen

Automatically scoring metaphor novelty is an unexplored topic in natural language processing, and research in this area could benefit a wide range of NLP tasks. However, no publicly available metaphor novelty datasets currently exist, making it difficult to perform research on this topic. We introduce a large corpus of metaphor novelty scores for syntactically related word pairs, and release it freely to the research community. We describe the corpus here, and include an analysis of its score distribution and the types of word pairs included in the corpus. We also provide a brief overview of standard metaphor detection corpora, to provide the reader with greater context regarding how this corpus compares to other datasets used for different types of computational metaphor processing. Finally, we establish a performance benchmark to which future researchers can compare, and show that it is possible to learn to score metaphor novelty on our dataset at a rate ignificantly better than chance or naive strategies.

Improving Hypernymy Extraction with Distributional Semantic Classes

Alexander Panchenko, Dmitry Ustalov, Stefano Faralli, Simone Paolo Ponzetto and Chris Biemann

In this paper, we show how distributionally-induced semantic classes can be helpful for extracting hypernyms. We present methods for inducing sense-aware semantic classes using distributional semantics and using these induced semantic classes for filtering noisy hypernymy relations. Denoising of hypernyms is performed by labeling each semantic class with its hypernyms. On the one hand, this allows us to filter out wrong extractions using the global structure of distributionally similar senses. On the other hand, we infer missing hypernyms via label propagation to cluster terms. We conduct a large-scale crowdsourcing study

showing that processing of automatically extracted hypernyms using our approach improves the quality of the hypernymy extraction in terms of both precision and recall. Furthermore, we show the utility of our method in the domain taxonomy induction task, achieving the state-of-the-art results on a SemEval'16 task on taxonomy induction.

Session O19 - Information Extraction & Neural Networks

10th May 2018, 11:45

Chair person: **Bernardo Magnini** Oral Session

Laying the Groundwork for Knowledge Base Population: Nine Years of Linguistic Resources for TAC KBP

Jeremy Getman, Joe Ellis, Stephanie Strassel, Zhiyi Song and Jennifer Tracey

Knowledge Base Population (KBP) is an evaluation series within the Text Analysis Conference (TAC) evaluation campaign conducted by the National Institute of Standards and Technology (NIST). Over the past nine years TAC KBP evaluations have targeted information extraction technologies for the population of knowledge bases comprised of entities, relations, and events. Linguistic Data Consortium (LDC) has supported TAC KBP since 2009, developing, maintaining, and distributing linguistic resources in three languages for seven distinct evaluation tracks. This paper describes LDC's resource creation efforts for the various KBP tracks, and highlights changes made over the years to support evolving evaluation requirements.

A Dataset for Inter-Sentence Relation Extraction using Distant Supervision

Angrosh Mandya, Danushka Bollegala, Frans Coenen and Katie Atkinson

This paper presents a benchmark dataset for the task of inter-sentence relation extraction. The paper explains the distant supervision method followed for creating the dataset for inter-sentence relation extraction, involving relations previously used for standard intra-sentence relation extraction task. The study evaluates baseline models such as bag-of-words and sequence based recurrent neural network models on the developed dataset and shows that recurrent neural network models as more useful for the task of intra-sentence relation extraction.Comparing the results of the present work on intra-sentence relation extraction with previous work on inter-sentence relation extraction, the study identifies the need for more sophisticated models to handle long-range information between entities across sentences.

Diacritics Restoration Using Neural Networks

Jakub Náplava, Milan Straka, Pavel Straňák and Jan Hajic

In this paper, we describe a novel combination of a character-level recurrent neural-network based model and a language model applied to diacritics restoration. In many cases in the past and still at present, people often replace characters with diacritics with their ASCII counterparts. Despite the fact that the resulting text is usually easy to understand for humans, it is much harder for further computational processing. This paper opens with a discussion of applicability of restoration of diacritics in selected languages. Next, we present a neural network-based approach to diacritics generation. The core component of our model is a bidirectional recurrent neural network operating at a character level. We evaluate the model on two existing datasets consisting of four European languages. When combined with a language model, our model reduces the error of current best systems by 20% to 64%. Finally, we propose a pipeline for obtaining consistent diacritics restoration datasets for twelve languages and evaluate our model on it. All the code is available under open source license on https://github.com/arahusky/diacritics_restoration.

Ensemble Romanian Dependency Parsing with Neural Networks

Radu Ion, Elena Irimia and Verginica Barbu Mititelu

SSPR (Semantics-driven Syntactic Parser for Romanian) is a neural network ensemble parser developed for Romanian (a Python 3.5 application based on the Microsoft Cognitive Toolkit 2.0 Python API) that combines the parsing decisions of a varying number (in our experiments, 3) of other parsers (MALT, RGB and MATE), using information from additional lexical, morpho-syntactic and semantic features. SSPR outperforms the best individual parser (MATE in our case) with 1.6% LAS points and it is in the same class with the top 5 Romanian performers at the CONLL 2017 dependency parsing shared task. The train and test sets were extracted from a Romanian dependency treebank we developed and validated in the Universal Dependencies format. The treebank, used in the CONLL 2017 Romanian track as well, is open licenced; the parser is available on request.

Session O20 - Dialogue, Emotion, Multimodality

10th May 2018, 11:45

Chair person: **Nick Campbell** Oral Session

Classifying Sluice Occurrences in Dialogue

Austin Baird, Anissa Hamza and Daniel Hardt

Ellipsis is an important challenge for natural language processing systems, and addressing that challenge requires large collections of relevant data. The dataset described by Anand and McCloskey (2015), consisting of 4100 occurrences, is an important step towards addressing this issue. However, many NLP technologies require much larger collections of data. Furthermore, previous collections of ellipsis are primarily restricted to news data, although sluicing presents a particularly important challenge for dialogue systems. In this paper we classify sluices as Direct, Reprise, Clarification. We perform manual annotation with acceptable inter-coder agreement. We build classifier models with Decision Trees and Naive Bayes, with accuracy of 67%. We deploy a classifier to automatically classify sluice occurrences in OpenSubtitles, resulting in a corpus with 1.7 million occurrences. This will support empirical research into sluicing in dialogue, and it will also make it possible to build NLP systems using very large datasets. This is a noisy dataset; based on a small manually annotated sample, we found that only 80% of instances are in fact sluices, and the accuracy of sluice classification is lower. Despite this, the corpus can be of great use in research on sluicing and development of systems, and we are making the corpus freely available on request. Furthermore, we are in the process of improving the accuracy of sluice identification and annotation for the purpose of created a subsequent version of this corpus.

Collection of Multimodal Dialog Data and Analysis of the Result of Annotation of Users' Interest Level

Masahiro Araki, Sayaka Tomimasu, Mikio Nakano, Kazunori Komatani, Shogo Okada, Shinya Fujie and Hiroaki Sugiyama

The Human-System Multimodal Dialogue Sharing Corpus Building Group is acting as a working group of SIG-SLUD for the purpose of constructing a corpus for evaluating elemental technologies of the multimodal dialogue system. In this paper, we report the results of recording chat dialogue data between a human and a virtual agent by the Wizard of OZ method conducted in 2016, and the results of the analysis of annotations of users' interest level in the data.

Recognizing Behavioral Factors while Driving: A Real-World Multimodal Corpus to Monitor the Driver's Affective State

Alicia Lotz, Klas Ihme, Audrey Charnoz, Pantelis Maroudis, Ivan Dmitriev and Andreas Wendemuth

The presented study concentrates on the collection of emotional multimodal real-world in-car audio, video and physiological signal recordings while driving. To do so, three sensor systems were integrated in the car and four relevant emotional states of the driver were defined: neutral, positive, frustrated and anxious. To gather as natural as possible emotional data of the driver, the subjects needed to be unbiased and were therefore kept unaware of the detailed research objective. The emotions were induced using so-called Wizard-of-Oz experiments, where the drivers believed to be interacting with an automated technical system, which in fact was controlled by a human. Additionally, on board interviews while driving were conducted by an instructed psychologist. To evaluate the collected data, questionnaires were filled out by the subjects before, during and after the data collection. These included monitoring of the drivers perceived state of emotion, stress, sleepiness and thermal sensation but also detailed questionnaires on their driving experience, attitude towards technology and big five OCEAN personality traits. Afterwards, the data was annotated by expert labelers. Exemplary results of the evaluation of the experiments are given in the result section of this paper. They indicate that the emotional states were successfully induced and the annotation results are consistent for both performed annotation approaches.

EmotionLines: An Emotion Corpus of Multi-Party Conversations

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang and Lun-Wei Ku

Feeling emotion is a critical characteristic to distinguish people from machines. Among all the multi-modal resources for emotion detection, textual datasets are those containing the least additional information in addition to semantics, and hence are adopted widely for testing the developed systems. However, most of the textual emotional datasets consist of emotion labels of only individual words, sentences or documents, which makes it challenging to discuss the contextual flow of emotions. In this paper, we introduce EmotionLines, the first dataset with emotions labeling on all utterances in each dialogue only based on their textual content. Dialogues in EmotionLines are collected from Friends TV scripts and private Facebook messenger dialogues. Then one of seven emotions, six Ekman's basic emotions plus the neutral emotion, is labeled on each utterance by 5 Amazon MTurkers. A total of 29,245 utterances from 2,000 dialogues are

labeled in EmotionLines. We also provide several strong baselines for emotion detection models on EmotionLines in this paper.

Session: I-P1 - Industry Track - Industrial Systems

10th May 2018, 11:45

Chair person: **Linne Ha** Poster Session

FonBund: A Library for Combining Cross-lingual Phonological Segment Data

Alexander Gutkin, Martin Jansche and Tatiana Merkulova

We present an open-source library (FonBund) that provides a way of mapping sequences of arbitrary phonetic segments in International Phonetic Alphabet (IPA) into multiple articulatory feature representations. The library interfaces with several existing linguistic typology resources providing phonological segment inventories and their corresponding articulatory feature systems. Our first goal was to facilitate the derivation of articulatory features without giving a special preference to any particular phonological segment inventory provided by freely available linguistic typology resources. The second goal was to build a very light-weight library that can be easily modified to support new phonological segment inventories. In order to support IPA segments that do not occur in the freely available resources, the library provides a simple configuration language for performing segment rewrites and adding custom segments with the corresponding feature structures. In addition to introducing the library and the corresponding linguistic resources, we also describe some of the practical uses of this library (multilingual speech synthesis) in the hope that this software will help facilitate multilingual speech research.

Voice Builder: A Tool for Building Text-To-Speech Voices

Pasindu De Silva, Theeraphol Wattanavekin, Tang Hao and Knot Pipatsrisawat

We describe an opensource text-to-speech (TTS) voice building tool that focuses on simplicity, flexibility, and collaboration. Our tool allows anyone with basic computer skills to run voice training experiments and listen to the resulting synthesized voice. We hope that this tool will reduce the barrier for creating new voices and accelerate TTS research, by making experimentation faster and interdisciplinary collaboration easier. We believe that our tool can help improve TTS research, especially for low-resourced languages, where more experimentations are often needed to get the most out of the limited language resources.

Sudachi: a Japanese Tokenizer for Business

Kazuma Takaoka, Sorami Hisamoto, Noriko Kawahara, Miho Sakamoto, Yoshitaka Uchida and Yuji Matsumoto

This paper presents Sudachi, a Japanese tokenizer and its accompanying language resources for business use. Tokenization, or morphological analysis, is a fundamental and important technology for processing a Japanese text, especially for industrial applications. However, we often face many obstacles for Japanese tokenization, such as the inconsistency of token unit in different resources, notation variations, discontinued maintenance of the resources, and various issues with the existing tokenizer implementations. In order to improve this situation, we develop a new tokenizer and a dictionary with features such as multi-granular output for different purposes and normalization of notation variations. In addition to this, we are planning to continuously maintain our software and resource in long-term as a part of the company business. We release the resulting tokenizer software and language resources freely available to the public as an open source software. You can access them at https://github.com/WorksApplications/Sudachi.

Chemical Compounds Knowledge Visualization with Natural Language Processing and Linked Data

Kazunari Tanaka, Tomoya Iwakura, Yusuke Koyanagi, Noriko Ikeda, Hiroyuki Shindo and Yuji Matsumoto

Knowledge of chemical compounds is invaliable for developing new materials, new drugs, and so on. Therefore, databases of chemical compounds are being created. For example, CAS, one of the largest databases, includes over 100 million chemical compound information. However, the creation of such databases strongly depends on manual labor since chemical compounds are being produced at every moment. In addition, the database creation mainly focuses on English text. Therefore, in other words, chemical compound information other than English is not good enough to be available. For example, although Japan has one of the largest chemical industries and has large chemical compound information written in Japanese text documents, such information is not exploited well so far. We propose a visualization system based on chemical compound extraction results with Japanese Natural Language Processing and structured databases represented as Linked Data (LD). Figure 1 shows an overview of our system. First, chemical compound names in text are recognized. Then, aliases of chemical compound names are identified. The extraction results and existing chemical compound databases are represented as LD. By combining these LD-based

chemical compound knowledge, our system provides different views of chemical compounds.

Session P26 - Language Acquisition & CALL (1)

10th May 2018, 11:45

Chair person: **Donghui Lin** Poster Session

Using Discourse Information for Education with a Spanish-Chinese Parallel Corpus

Shuyuan Cao and Harritxu Gete

Nowadays, with the fruitful achievements in Natural Language Processing (NLP) studies, the concern of using NLP technologies for education has called much attention. As two of the most spoken languages in the world, Spanish and Chinese occupy important positions in both NLP studies and bilingual education. In this paper, we present a Spanish-Chinese parallel corpus with annotated discourse information that aims to serve for bilingual language education. The theoretical framework of this work is Rhetorical Structure Theory (RST). The corpus is composed of 100 Spanish-Chinese parallel texts, and all the discourse markers (DM) have been annotated to form the education source. With pedagogical aim, we also present two programs that generate automatic exercises for both Spanish and Chinese students using our corpus. The reliability of this work has been evaluated using Kappa coefficient.

A 2nd Longitudinal Corpus for Children's Writing with Enhanced Output for Specific Spelling Patterns

Kay Berkling

This paper describes the collection of three longitudinal Corpora of German school children's weekly writing in German, called H2 (H1 is available via LDC and contains some of the same students' writing 2 years previously), E2 (E1 is not public), and ERK1. The texts were written within the normal classroom setting. Texts of children whose parents signed the permission to donate the texts to science were collected and transcribed. The corpus consists of the elicitation techniques, an overview of the data collected and the transcriptions of the texts both with and without spelling errors, aligned on a word by word basis. In addition, the hand-written texts were scanned in. The corpus is available for research via Linguistic Data Consortium (LDC). When using this Corpus, researchers are strongly encouraged to make additional annotations and improvements and return it to the public domain via LDC, especially since this effort was unfunded.

Development of a Mobile Observation Support System for Students: FishWatchr Mini

Masaya Yamaguchi, Masanori Kitamura and Naomi Yanagida

This paper presents a system called FishWatchr Mini (FWM), which supports students in observing and reflecting on educational activities such as presentation practices. Although video annotation systems are supposed to support such activities, especially reflection, they have been used by researchers and teachers outside the classroom, rather than by students inside the classroom. To resolve problems with introducing video annotation systems in the classroom, we propose three solutions: (a) to facilitate preparation of devices, FWM is implemented as a Web application on students' smartphones; (b) to facilitate students' learning to use the system, FWM is designed to automate as many operations as possible, apart from annotation; and (c) FWM allows students to examine annotation data through reflection, by providing functions such as visualization. The results of applying FWM to presentation practices in a university validated the effectiveness of FWM in terms of its ease of use and applicability of annotation results. Since annotated video data could also provide language resources for teachers, it is anticipated that they will contribute to improving their classes in the future.

The AnnCor CHILDES Treebank

Jan Odijk, Alexis Dimitriadis, Martijn Van der Klis, Marjo Van Koppen, Meie Otten and Remco Van der Veen

This paper (1) presents the first partially manually verified treebank for Dutch CHILDES corpora, the AnnCor CHILDES Treebank; (2) argues explicitly that it is useful to assign adult grammar syntactic structures to utterances of children who are still in the process of acquiring the language; (3) argues that human annotation and automatic checks on this annotation must go hand in hand; (4) argues that explicit annotation guidelines and conventions must be developed and adhered to and emphasises consistency of the annotations as an important desirable property for annotations. It also describes the tools used for annotation and automated checks on edited syntactic structures, as well as extensions to an existing treebank query application (GrETEL) and the multiple formats in which the resources will be made available.

BabyCloud, a Technological Platform for Parents and Researchers

Xuan-Nga Cao, Cyrille Dakhlia, Patricia Del Carmen, Mohamed-Amine Jaouani, Malik Ould-Arbi and Emmanuel Dupoux

In this paper, we present BabyCloud, a platform for capturing, storing and analyzing daylong audio recordings and photographs

of children's linguistic environments, for the purpose of studying infant's cognitive and linguistic development and interactions with the environment. The proposed platform connects two communities of users: families and academics, with strong innovation potential for each type of users. For families, the platform offers a novel functionality: the ability for parents to follow the development of their child on a daily basis through language and cognitive metrics (growth curves in number of words, verbal complexity, social skills, etc). For academic research, the platform provides a novel means for studying language and cognitive development at an unprecedented scale and level of detail. They will submit algorithms to the secure server which will only output anonymized aggregate statistics. Ultimately, {BabyCloud aims at creating an ecosystem of third parties (public and private research labs...) gravitating around developmental data, entirely controlled by the party whose data originate from, i.e. families.

Infant Word Comprehension-to-Production Index Applied to Investigation of Noun Learning Predominance Using Cross-lingual CDI database

Yasuhiro Minami, Tessei Kobayashi and Yuko Okumura

This paper defines a measure called the comprehension-to-production (C2P) index to investigate whether nouns have predominance over verbs in children's word learning that identifies a partial word learning period from comprehension to production. We applied the C2P index to noun predominance using cross-lingual child communicative development inventory databases and confirmed that it indicates noun predominance in word learning, suggesting that the process between a word's comprehension and its production is a significant factor of predominance in noun learning by children.

Building a TOCFL Learner Corpus for Chinese Grammatical Error Diagnosis

Lung-Hao Lee, Yuen-Hsien Tseng and Liping Chang

This study describes the construction of a TOCFL learner corpus and its usage for Chinese grammatical error diagnosis. We collected essays from the Test Of Chinese as a Foreign Language (TOCFL) and annotated grammatical errors using hierarchical tagging sets. Two kinds of error classifications were used simultaneously to tag grammatical errors. The first capital letter of each error tags denotes the coarse-grained surface differences, while the subsequent lowercase letters denote the fine-grained linguistic categories. A total of 33,835 grammatical errors in 2,837 essays and their corresponding corrections were manually annotated. We then used the Standard Generalized Markup Language to format learner texts and annotations along with learners' accompanying metadata. Parts of the TOCFL learner corpus have been provided for shared tasks on Chinese grammatical error diagnosis. We also investigated systems participating in the shared tasks to better understand current achievements and challenges. The datasets are publicly available to facilitate further research. To our best knowledge, this is the first annotated learner corpus of traditional Chinese, and the entire learner corpus will be publicly released.

MIAPARLE: Online training for the discrimination of stress contrasts

Jean-Philippe Goldman and Sandra Schwab

MIAPARLE is a public web application that is designed to offer a range of CAPT (computer-aided pronunciation teaching) tools for L2 learners. Besides helping language learners to reduce their foreign-accentedness, the goal of the platform is to test these tools, gather feedback and improve them according to their educational impact. In this article, we describe one particular training tool that focuses on stress perception. This tool is particularly useful for speakers whose L1 is a fixed-stress language, such as French. These speakers have difficulties perceiving and discriminating stress contrasts. To help them with this so-called stress 'deafness', the methodology used in the training is based on successive questions in which a visual pattern is associated with the sound of a lexical item. After successively completing their pre-tests, training and post-tests, the participants are given their improvement score. The performance of the training is evaluated by comparing the learner's results at the pre- and post-test stages. Various methodological parameters, such as the number of training items or the number of visual patterns are tested in parallel in order to quantify their teaching efficiency, and to optimise the overall teaching impact.

ESCRITO - An NLP-Enhanced Educational Scoring Toolkit

Torsten Zesch and Andrea Horbach

We propose Escrito, a toolkit for scoring student writings using NLP techniques that addresses two main user groups: teachers and NLP researchers. Teachers can use a high-level API in the teacher mode to assemble scoring pipelines easily. NLP researchers can use the developer mode to access a low-level API, which not only makes available a number of pre-implemented components, but also allows the user to integrate their own readers, preprocessing components, or feature extractors. In this way, the toolkit provides a ready-made testbed for applying the latest developments from NLP areas like text similarity, paraphrase detection, textual entailment, and argument mining within the highly challenging task of educational scoring and feedback. At the same time, it allows teachers to apply cutting-edge technology in the classroom.

A Leveled Reading Corpus of Modern Standard Arabic

Muhamed Al Khalil, Hind Saddiki, Nizar Habash and Latifa Alfalasi

We present a reading corpus in Modern Standard Arabic to enrich the sparse collection of resources that can be leveraged for educational applications. The corpus consists of textbook material from the curriculum of the United Arab Emirates, spanning all 12 grades (1.4 million tokens) and a collection of 129 unabridged works of fiction (5.6 million tokens) all annotated with reading levels from Grade 1 to Post-secondary. We examine reading progression in terms of lexical coverage, and compare the two sub-corpora (curricular, fiction) to others from clearly established genres (news, legal/diplomatic) to measure representation of their respective genres.

Session P27 - Less-Resourced/Endangered Languages (1)

10th May 2018, 11:45

Chair person: **Valérie Mapelli** Poster Session

Developing New Linguistic Resources and Tools for the Galician Language

Rodrigo Agerri, Xavier Gómez Guinovart, German Rigau and Miguel Anxo Solla Portela

In this paper we describe the work towards developing new resources and Natural Language Processing (NLP) tools for the Galician language. First, a new corpus, manually revised, for POS tagging and lemmatization is described. Second, we present a new manually annotated corpus for Named Entity tagging for Galician. Third, we train and develop new NLP tools for Galician, including the first publicly available Galician statistical modules for lemmatization and Named Entity Recognition, and new modules for POS tagging, Wikification and Named Entity Disambiguation. Finally, we also present two new Web demo applications to easily test the new set of tools online.

Modeling Northern Haida Verb Morphology

Jordan Lachler, Lene Antonsen, Trond Trosterud, Sjur Moshagen and Antti Arppe

This paper describes the development of a computational model of the morphology of Northern Haida based on finite state machines (FSMs), with a focus on verbs. Northern Haida is highly endangered, and a member of the isolate Haida macrolanguage, spoken in British Columbia and Alaska. Northern Haida is a highly-inflecting language whose verbal morphology relies largely on suffixes, with a limited number of prefixes. The suffixes trigger morphophonological changes in the stem, participate in blocking, and exhibit variable ordering in certain constructions. The computational model of Northern Haida verb morphology is capable of handling these complex affixation patterns and the morphophonological alternations that they engender. In this paper, we describe the challenges we encountered and the solutions we propose, while contextualizing the endeavour in the description, documentation and revitalization of First Nations Languages in Canada.

Low-resource Post Processing of Noisy OCR Output for Historical Corpus Digitisation

Caitlin Richter, Matthew Wickes, Deniz Beser and Mitchell Marcus

We develop a post-processing system to efficiently correct errors from noisy optical character recognition (OCR) in a 2.7 million word Faroese corpus. 7.6% of the words in the original OCR text contain an error; fully manual correction would take thousands of hours due to the size of the corpus. Instead, our post-processing method applied to the Faroese corpus is projected to reduce the word error rate to 1.3% with around 65 hours of human annotator work. The foundation for generating corrected text is an HMM that learns patterns of OCR error and decodes noisy OCR character sequences into hypothesised correct language. A dictionary augments the HMM by contributing additional language knowledge, and a human annotator provides judgements in a small subset of cases that are identified as otherwise most prone to inaccurate output. An interactive workstation facilitates quick and accurate input for annotation. The entire toolkit is written in Python and is being made available for use in other low-resource languages where standard OCR technology falls short of desirable text quality. Supplementary analyses explore the impact of variable language resource availability and annotator time limitations on the end quality achievable with our toolkit.

Introducing the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation

Hanna Hedeland, Timm Lehmberg, Felix Rau, Sophie Salffner, Mandana Seyfeddinipur and Andreas Witt

The European digital research infrastructure CLARIN (Common Language Resources and Technology Infrastructure) is building a Knowledge Sharing Infrastructure (KSI) to ensure that existing knowledge and expertise is easily available both for the CLARIN community and for the humanities research communities for which CLARIN is being developed. Within the Knowledge Sharing Infrastructure, so called Knowledge Centres comprise one or more physical institutions with particular expertise in

certain areas and are committed to providing their expertise in the form of reliable knowledge-sharing services. In this paper, we present the ninth K Centre – the CLARIN Knowledge Centre for Linguistic Diversity and Language Documentation (CKLD) – and the expertise and services provided by the member institutions at the Universities of London (ELAR/SWLI), Cologne (DCH/IfDH/IfL) and Hamburg (HZSK/INEL). The centre offers information on current best practices, available resources and tools, and gives advice on technological and methodological matters for researchers working within relevant fields.

Low Resource Methods for Medieval Document Sections Analysis

Petra Galuscakova and Lucie Neuzilova

This paper describes a small but unique digitized collection of medieval Latin charters. This collection consists of 57 charters of 7 types illustrating various purposes of issuance by the Royal Chancellery. Sections in these documents were manually annotated for deeper analysis of the structure of issued charters. This paper also describes two baseline methods for an automatic and semi-automatic analysis and detection of sections of diplomatic documents. The first method is based on an information retrieval paradigm, and the second one is an adaptation of Hidden Markov Models. Both methods were proposed to work with respect to a small amount of available train data. Even though these methods were specifically proposed to work with medieval Latin charters, they can be applied to any documents with partially repetitive character.

SB-CH: A Swiss German Corpus with Sentiment Annotations

Ralf Grubenmann, Don Tuggener, Pius Von Däniken, Jan Deriu and Mark Cieliebak

We present the SB-CH corpus, a novel Swiss German corpus with annotations for sentiment analysis. It consists of more than 200,000 phrases (approx. 1 Mio tokens) from Facebook comments and online chats. Additionally, we provide sentiment annotations for almost 2000 Swiss German phrases. We describe the methodologies used in the collection and annotation of the data, and provide the first baseline results for Swiss German sentiment analysis.

Universal Dependencies for Ainu

Hajime Senuma and Akiko Aizawa

This paper reports an on-going effort to create a dependency tree bank for the Ainu language in the scheme of Universal Dependencies (UD). The task is crucial both language-internally (language revitalization) and language-externally (providing sources for new features and insights to UD). Since the language shows many of the representative phenomena of a type of languages called polysynthetic languages, an annotation schema to Ainu can be used as a basis to extend the current specification of UD. Our language resource comprises an annotation guideline, dependency bank based on UD, and a mini-lexicon. Although the size of the dependency bank will be small and contain only around 10,000 word tokens, it can serve as a base annotation for the next step. Our mini-lexicon is encoded under the W3C OntoLex specification with UD and UniMorph (UM) features with the system-friendly JSON-LD format and thus bearable to future extensions. We also provide a brief description of dependency relations and local features used in the bank such as pronominal cross-indexing and alienability.

Session P28 - Lexicon (2)

10th May 2018, 11:45

Chair person: **Lionel Nicolas** Poster Session

Signbank: Software to Support Web Based Dictionaries of Sign Language

Steve Cassidy, Onno Crasborn, Henri Nieminen, Wessel Stoop, Micha Hulsbosch, Susan Even, Erwin Komen and Trevor Johnson

Signbank is a web application that was originally built to support the Auslan Signbank on-line web dictionary, it was an Open Source re-implementation of an earlier version of that site. The application provides a framework for the development of a rich lexical database of sign language augmented with video samples of signs. As an Open Source project, the original Signbank has formed the basis of a number of new sign language dictionaries and corpora including those for British Sign Language, Sign Language of the Netherlands and Finnish Sign Language. Versions are under development for American Sign Language and Flemish Sign Language. This paper describes the overall architecture of the Signbank system and its representation of lexical entries and associated entities.

J-MeDic: A Japanese Disease Name Dictionary based on Real Clinical Usage

Kaoru Ito, Hiroyuki Nagai, Taro Okahisa, Shoko Wakamiya, Tomohide Iwao and Eiji Aramaki

Medical texts such as electronic health records are necessary for medical AI development. Nevertheless, it is difficult to use data directly because medical texts are written mostly in natural language, requiring natural language processing (NLP) for medical texts. To boost the fundamental accuracy of Medical NLP, a high coverage dictionary is required, especially one that fills the gap separating standard medical names and real clinical words.

This study developed a Japanese disease name dictionary called "J-MeDic" to fill this gap. The names that comprise the dictionary were collected from approximately 45,000 manually annotated real clinical case reports. We allocated the standard disease code (ICD-10) to them with manual, semi-automatic, or automatic methods, in accordance with its frequency. The J-MeDic covers 7,683 concepts (in ICD-10) and 51,784 written forms. Among the names covered by J-MeDic, 55.3% (6,391/11,562) were covered by SDNs; 44.7% (5,171/11,562) were covered by names added from the CR corpus. Among them, 8.4% (436/5,171) were basically coded by humans), and 91.6% (4,735/5,171) were basically coded automatically. We investigated the coverage of this resource using discharge summaries from a hospital; 66.2% of the names are matched with the entries, revealing the practical feasibility of our dictionary.

Building a List of Synonymous Words and Phrases of Japanese Compound Verbs

Kyoko Kanzaki and Hitoshi Isahara

We started to construct a database of synonymous expressions of Japanese "Verb + Verb" compounds semi-automatically. Japanese is known to be rich in compound verbs consisting of two verbs joined together. However, we did not have a comprehensive Japanese compound lexicon. Recently a Japanese compound verb lexicon was constructed by the National Institute for Japanese Language and Linguistics(NINJAL)(2013-15). Though it has meanings, example sentences, syntactic patterns and actual sentences from the corpus that they possess, it has no information on relationships with another words, such as synonymous words and phrases. We automatically extracted synonymous expressions of compound verbs from corpus which is "five hundred million Japanese texts gathered from the web" produced by Kawahara et.al. (2006) by using word2vec and cosine similarity and find suitable clusters which correspond to meanings of the compound verbs by using k-means++ and PCA. The automatic extraction from corpus helps humans find not only typical synonyms but also unexpected synonymous words and phrases. Then we manually compile the list of synonymous expressions of Japanese compound verbs by assessing the result and also link it to the "Compound Verb Lexicon" published by NINJAL.

Evaluating EcoLexiCAT: a Terminology-Enhanced CAT Tool

Pilar León-Araúz and Arianne Reimerink

EcoLexiCAT is a web-based tool for the terminology-enhanced translation of specialized environmental texts for the language combination English-Spanish-English. It uses the open source version of the web-based CAT tool MateCat and enriches a source text with information from: (1) EcoLexicon, a multimodal and multilingual terminological knowledge base on the environment (Faber et al., 2014; Faber et al., 2016); (2) BabelNet, an automatically constructed multilingual encyclopedic dictionary and semantic network (Navigli & Ponzetto, 2012); (3) Sketch Engine, the well-known corpus query system (Kilgarriff et al., 2004); (4) IATE, the multilingual glossary of the European Commission; and (4) other external resources (i.e. Wikipedia, Collins, Wordreference, Linguee, etc.) that can also be customized by the user. The tool was built with the aim of integrating terminology management – often considered complex and time-consuming – in the translation workflow of a CAT tool. In this paper, EcoLexiCAT is described along the procedure with which it was evaluated and the results of the evaluation.

A Danish FrameNet Lexicon and an Annotated Corpus Used for Training and Evaluating a Semantic Frame Classifier

Bolette Pedersen, Sanni Nimb, Anders Søgaard, Mareike Hartmann and Sussi Olsen

In this paper, we present an approach to efficiently compile a Danish FrameNet based on the Danish Thesaurus, focusing in particular on cognition and communication frames. The Danish FrameNet uses the frame and role inventory of the English FrameNet. We present the corresponding corpus annotations of frames and roles and show how our corpus can be used for training and evaluating a semantic frame classifier for cognition and communication frames. We also present results of cross-language transfer of a model trained on the English FrameNet. Our approach is significantly faster than building a lexicon from scratch, and we show that it is feasible to annotate Danish with frames developed for English, and finally, that frame annotations – even if limited in size at the current stage – are useful for automatic frame classification.

SLIDE - a Sentiment Lexicon of Common Idioms

Charles Jochim, Francesca Bonin, Roy Bar-Haim and Noam Slonim

Idiomatic expressions are problematic for most sentiment analysis approaches, which rely on words as the basic linguistic unit. Compositional solutions for phrase sentiment are not able to handle idioms correctly because their sentiment is not derived from the sentiment of the individual words. Previous work has explored the importance of idioms for sentiment analysis, but has not addressed the breadth of idiomatic expressions in English. In this paper we present an approach for collecting sentiment annotation of idiomatic multiword expressions using crowdsourcing. We collect 10 annotations for each idiom and the aggregated label is shown to have good agreement with expert annotations. We describe the resulting publicly available lexicon

and how it captures sentiment strength and ambiguity. The Sentiment Lexicon of IDiomatic Expressions (SLIDE) is much larger than previous idiom lexicons. The lexicon includes 5,000 frequently occurring idioms, as estimated from a large English corpus. The idioms were selected from Wiktionary, and over 40% of them were labeled as sentiment-bearing.

PronouncUR: An Urdu Pronunciation Lexicon Generator

Haris Bin Zia, Agha Ali Raza and Awais Athar

State-of-the-art speech recognition systems rely heavily on three basic components: an acoustic model, a pronunciation lexicon and a language model. To build these components, a researcher needs linguistic as well as technical expertise, which is a barrier in low-resource domains. Techniques to construct these three components without having expert domain knowledge are in great demand. Urdu, despite having millions of speakers all over the world, is a low-resource language in terms of standard publically available linguistic resources. In this paper, we present a grapheme-to-phoneme conversion tool for Urdu that generates a pronunciation lexicon in a form suitable for use with speech recognition systems from a list of Urdu words. The tool predicts the pronunciation of words using a LSTM-based model trained on a handcrafted expert lexicon of around 39,000 words and shows an accuracy of 64% upon internal evaluation. For external evaluation on a speech recognition task, we obtain a word error rate comparable to one achieved using a fully handcrafted expert lexicon.

SimLex-999 for Polish

Agnieszka Mykowiecka, Malgorzata Marciniak and Piotr Rychlik

The paper addresses the Polish version of SimLex-999 which we extended to contain not only measurement of similarity but also relatedness. The data was translated by three independent linguists; discrepancies in translation were resolved by a fourth person. The agreement rates between the translators were counted and an analysis of problems was performed. Then, pairs of words were rated by other annotators on a scale of 0–10 for similarity and relatedness of words. Finally, we compared the human annotations with the distributional semantics models of Polish based on lemmas and forms. We compared our work with the results reported for other languages.

Finely Tuned, 2 Billion Token Based Word Embeddings for Portuguese

João Rodrigues and António Branco

A distributional semantics model — also known as word embeddings — is a major asset for any language as the research results reported in the literature have consistently shown that it is instrumental to improve the performance of a wide range of applications and processing tasks for that language. In this paper, we describe the development of an advanced distributional model for Portuguese, with the largest vocabulary and the best evaluation scores published so far. This model was made possible by resorting to new languages resources we recently developed: to a much larger training corpus than before and to a more sophisticated evaluation supported by new and more fine-grained evaluation tasks and data sets. We also indicate how the new language resource reported on here is being distributed and where it can be obtained for free under a most permissive license.

Session P29 - Linked Data

10th May 2018, 11:45

Chair person: **Monica Monachini** Poster Session

Teanga: A Linked Data based platform for Natural Language Processing

Housam Ziad, John Philip McCrae and Paul Buitelaar

In this paper, we describe Teanga, a linked data based platform for natural language processing (NLP). Teanga enables the use of many NLP services from a single interface, whether the need was to use a single service or multiple services in a pipeline. Teanga focuses on the problem of NLP services interoperability by using linked data to define the types of services input and output. Teanga's strengths include being easy to install and run, easy to use, able to run multiple NLP tasks from one interface and helping users to build a pipeline of tasks through a graphical user interface.

Automatic and Manual Web Annotations in an Infrastructure to handle Fake News and other Online Media Phenomena

Georg Rehm, Julian Moreno-Schneider and Peter Bourgonje

Online media are ubiquitous and consumed by billions of people globally. Recently, however, several phenomena regarding online media have emerged that pose a severe threat to media consumption and reception as well as to the potential of manipulating opinions and, thus, (re)actions, on a large scale. Lumped together under the label "fake news", these phenomena comprise, among others, maliciously manipulated content, bad journalism, parodies, satire, propaganda and several other types of false news; related phenomena are the often cited filter bubble (echo chamber) effect and the amount of abusive language used online. In an earlier paper we describe an architectural and technological approach to empower users to handle these online

media phenomena. In this article we provide the first approach of a metadata scheme to enable, eventually, the standardised annotation of these phenomena in online media. We also show an initial version of a tool that enables the creation, visualisation and exploitation of such annotations.

The LODeXporter: Flexible Generation of Linked Open Data Triples from NLP Frameworks for Automatic Knowledge Base Construction

René Witte and Bahar Sateli

We present LODeXporter, a novel approach for exporting Natural Language Processing (NLP) results to a graph-based knowledge base, following Linked Open Data (LOD) principles. The rules for transforming NLP entities into Resource Description Framework (RDF) triples are described in a custom mapping language, which is defined in RDF Schema (RDFS) itself, providing a separation of concerns between NLP pipeline engineering and knowledge base engineering. LODeXporter is available as an open source component for the GATE (General Architecture for Text Engineering) framework.

LiDo RDF: From a Relational Database to a Linked Data Graph of Linguistic Terms and Bibliographic Data

Bettina Klimek, Robert Schädlich, Dustin Kröger, Edwin Knese and Benedikt Elßmann

Forty years ago the linguist Dr. Christian Lehmann developed a framework for documenting linguistic terms, concepts and bibliographic data that resulted in the LiDo Terminological and Bibliographical Database (LiDo TBD). Since 2006 students and linguistic researchers benefit from the data by looking it up on the Web. Even though, the LiDo TBD is implemented as a relational database, its underlying framework aims at yielding a terminological network containing data nodes that are connected via specific relation edges in order to create an interrelated data graph. Now, with the emergence of Semantic Web technologies we were able to implement this pioneering work by converting the LiDo TBD relational database into a Linked Data graph. In this paper we present and describe the creation of the LiDo RDF dataset and introduce the LiDo RDF project. The goals of this project are to enable the direct use and reuse of the data both for the scientific research community and machine processing alike as well as to enable a valuable enrichment of already existing linguistic terminological and bibliographic data by including LiDo RDF in the LLOD cloud.

Towards a Linked Open Data Edition of Sumerian Corpora

Christian Chiarcos, Émilie Pagé-Perron, Ilya Khait, Niko Schenk and Lucas Reckling

Linguistic Linked Open Data (LLOD) is a flourishing line of research in the language resource community, so far mostly adopted for selected aspects of linguistics, natural language processing and the semantic web, as well as for practical applications in localization and lexicography. Yet, computational philology seems to be somewhat decoupled from the recent progress in this area: even though LOD as a concept is gaining significant popularity in Digital Humanities, existing LLOD standards and vocabularies are not widely used in this community, and philological resources are underrepresented in the LLOD cloud diagram (http://linguistic-lod.org/llod-cloud). In this paper, we present an application of Linguistic Linked Open Data in Assyriology. We describe the LLOD edition of a linguistically annotated corpus of Sumerian, as well as its linking with lexical resources, repositories of annotation terminology, and the museum collections in which the artifacts bearing these texts are kept. The chosen corpus is the Electronic Text Corpus of Sumerian Royal Inscriptions, a well curated and linguistically annotated archive of Sumerian text, in preparation for the creating and linking of other corpora of cuneiform texts, such as the corpus of Ur III administrative and legal Sumerian texts, as part of the Machine Translation and Automated Analysis of Cuneiform Languages project(https://cdli-gh.github.io/mtaac/).

Session P30 - Infrastructural Issues/Large Projects (2)

10th May 2018, 11:45

Chair person: **Krister Lindén** Poster Session

A Bird's-eye View of Language Processing Projects at the Romanian Academy

Dan Tufiș and Cristea Dan

This article gives a general overview of five AI language-related projects that address contemporary Romanian language, in both textual and speech form, language related applications, as well as collections of old historic documents and medical archives. Namely, these projects deal with: the creation of a contemporary Romanian language text and speech corpus, resources and technologies for developing human-machine interfaces in spoken Romanian, digitization and transcription of old Romanian language documents drafted in Cyrillic into the modern Latin alphabet, digitization of the oldest archive of diabetes medical records and dialogue systems with personal

robots and autonomous vehicles. The technologies involved for attaining the objectives range from image processing (intelligent character recognition for hand-writing and old Romanian documents) to natural language and speech processing techniques (corpus compiling and documentation, multi-level processing, transliteration of different old scripts into modern Romanian, command language processing, various levels of speech-text alignments, ASR, TTS, keyword spotting, etc.). Some of these projects are approaching the end, others have just started and others are about to start. All the reported projects are national ones, less documented than the international projects we are/were engaged in, and involve large teams of experts and master/PhD students from computer science, mathematics, linguistics, philology and library sciences.

PMKI: an European Commission action for the interoperability, maintainability and sustainability of Language Resources

Peter Schmitz, Enrico Francesconi, Najeh Hajlaoui and Brahim Batouche

The paper presents the Public Multilingual Knowledge Management Infrastructure (PMKI) action launched by the European Commission (EC) to promote the Digital Single Market in the European Union (EU). PMKI aims to share maintainable and sustainable Language Resources making them interoperable in order to support language technology industry, and public administrations, with multilingual tools able to improve cross border accessibility of digital services. The paper focuses on the main feature (interoperability) that represents the specificity of PMKI platform distinguishing it from other existing frameworks. In particular it aims to create a set of tools and facilities, based on Semantic Web technologies, to establish semantic interoperability between multilingual lexicons. Such task requires to harmonize in general multilingual language resources using standardised representation with respect to a defined core data model under an adequate architecture. A comparative study among the main data models for representing lexicons and recommendations for the PMKI service was required as well. Moreover, synergies with other programs of the EU institutions, as far as systems interoperability and Machine Translation (MT) solutions, are foreseen. For instance some interactions are foreseen between PMKI and MT service provided by the EC but also with other NLP applications.

The Abkhaz National Corpus

Paul Meurer

In this paper, we present the Abkhaz National Corpus, a comprehensive and open, grammatically annotated text corpus which makes the Abkhaz language accessible to scientific investigations from various perspectives (linguistics, literary studies, history, political and social sciences etc.). The corpus also serves as a means for the long-term preservation of Abkhaz language documents in digital form, and as a pedagogical tool for language learning. It now comprises more than 10 million words and is continuously being extended. Abkhaz is a lesser-resourced language; prior to this work virtually no computational resources for the language were available. As a member of the West-Caucasian language family, which is characterized by an extremely rich, polysynthetic morphological structure, Abkhaz poses serious challenges to morphosyntactic analysis, the main problem being the high degree of morphological ambiguity. We show how these challenges can be met, and what we plan to further enhance the performance of the analyser.

Collecting Language Resources from Public Administrations in the Nordic and Baltic Countries

Andrejs Vasiljevs, Rihards Kalniņš, Roberts Rozis and Aivars Bērziņš

This paper presents Tilde's work on collecting language resources from government institutions and other public administrations in the Nordic and Baltic countries. We introduce the activities and results of the European Language Resources Coordination (ELRC) action in this region, provide a synopsis of ELRC workshops held in all countries of the region, identify potential holders and donors of language data suitable for improving machine translation systems, and describe the language resources collected so far. We also describe several national projects and initiatives on sharing of language data accumulated in the public sector and creation of new language resources from this data. Opportunities and challenges in consolidating language data from the public sector are discussed, and related actions and regulatory initiatives are proposed.

Session P31 - MultiWord Expressions & Collocations

10th May 2018, 11:45

Chair person: **Brigitte Krenn** Poster Session

LIdioms: A Multilingual Linked Idioms Data Set

Diego Moussallem, Mohamed Ahmed Sherif, Diego Esteves, Marcos Zampieri and Axel-Cyrille Ngonga Ngomo

In this paper, we describe the LIDIOMS data set, a multilingual RDF representation of idioms currently containing five languages:

English, German, Italian, Portuguese, and Russian. The data set is intended to support natural language processing applications by providing links between idioms across languages. The underlying data was crawled and integrated from various sources. To ensure the quality of the crawled data, all idioms were evaluated by at least two native speakers. Herein, we present the model devised for structuring the data. We also provide the details of linking LIDIOMS to well-known multilingual data sets such as BabelNet. The resulting data set complies with best practices according to Linguistic Linked Open Data Community.

Annotating Modality Expressions and Event Factuality for a Japanese Chess Commentary Corpus

Suguru Matsuyoshi, Hirotaka Kameko, Yugo Murawaki and Shinsuke Mori

In recent years, there has been a surge of interest in the natural language processing related to the real world, such as symbol grounding, language generation, and nonlinguistic data search by natural language queries. We argue that shogi (Japanese chess) commentaries, which are accompanied by game states, are an interesting testbed for these tasks. A commentator refers not only to the current board state but to past and future moves, and yet such references can be grounded in the game tree, possibly with the help of modern game-tree search algorithms. For this reason, we previously collected shogi commentaries together with board states and have been developing a game commentary generator. In this paper, we augment the corpus with manual annotation of modality expressions and event factuality. The annotated corpus includes 1,622 modality expressions, 5,014 event class tags and 3,092 factuality tags. It can be used to train a computer to identify words and phrases that signal factuality and to determine events with the said factuality, paving the way for grounding possible and counterfactual states.

Annotating Chinese Light Verb Constructions according to PARSEME guidelines

Menghan Jiang, Natalia Klyueva, Hongzhi Xu and Chu-Ren Huang

In this paper we present a preliminary study on application of PARSEME guidelines of annotating mutiword expressions to Chinese language. We focus on one specific category - light verb constructions (LVCs). We make use of an existing resource containing Chinese light verbs and examine whether this resource fulfill the requirements of the guidelines. We make a preliminary annotation of a Chinese UD treebank in two steps: first automatically identifying potential light verbs and then manually assigning the corresponding nouns or correcting false positives.

Using English Baits to Catch Serbian Multi-Word Terminology

Cvetana Krstev, Branislava Šandrih, Ranka Stankovic and Miljana Mladenović

In this paper we present the first results in bilingual terminology extraction. The hypothesis of our approach is that if for a source language domain terminology exists as well as a domain aligned corpus for a source and a target language, then it is possible to extract the terminology for a target language. Our approach relies on several resources and tools: aligned domain texts, domain terminology for a source language, a terminology extractor for a target language, and a tool for word and chunk alignment. In this first experiment a source language is English, a target language is Serbian, a domain is Library and Information Science for which a bilingual terminological dictionary exists. Our term extractor is based on e-dictionaries and shallow parsing, and for word alignment we use GIZA++. At the end of procedure we included a supervised binary classifier that decides whether an extracted term is a valid domain term. The classifier was evaluated in a 5-fold cross validation setting on a slightly unbalanced dataset, maintaining average F-score of 89%. After conducting the experiment our system extracted 846 different Serbian domain phrases, containing 515 Serbian phrases that were not present in the existing domain terminology.

Construction of Large-scale English Verbal Multiword Expression Annotated Corpus

Akihiko Kato, Hiroyuki Shindo and Yuji Matsumoto

Multiword expressions (MWEs) consist of groups of tokens, which should be treated as a single syntactic or semantic unit. In this work, we focus on verbal MWEs (VMWEs), whose accurate recognition is challenging because they could be discontinuous (e.g., take .. off). Since previous English VMWE annotations are relatively small-scale in terms of VMWE occurrences and types, we conduct large-scale annotations of VMWEs on the Wall Street Journal portion of English Ontonotes by a combination of automatic annotations and crowdsourcing. Concretely, we first construct a VMWE dictionary based on the English-language Wiktionary. After that, we collect possible VMWE occurrences in Ontonotes and filter candidates with the help of gold dependency trees, then we formalize VMWE annotations as a multiword sense disambiguation problem to exploit crowdsourcing. As a result, we annotate 7,833 VMWE instances belonging to various categories, such as phrasal verbs, light verb constructions, and semi-fixed VMWEs. We hope this large-scale VMWE-annotated resource helps to develop models for MWE recognition and dependency parsing that are aware of English MWEs. Our resource is publicly available.

Konbitzul: an MWE-specific database for Spanish-Basque

Uxoa Iñurrieta, Itziar Aduriz, Arantza Diaz de Ilarraza, Gorka Labaka and Kepa Sarasola

This paper presents Konbitzul, an online database of verb+noun MWEs in Spanish and Basque. It collects a list of MWEs with their translations, as well as linguistic information which is NLP-applicable: it helps to identify occurrences of MWEs in multiple morphosyn- tactic variants, and it is also useful for improving translation quality in rule-based MT. In addition to this, its user-friendly interface makes it possible to simply search for MWEs along with translations, just as in any bilingual phraseological dictionary.

A Multilingual Test Collection for the Semantic Search of Entity Categories

Juliano Efson Sales, Siamak Barzegar, Wellington Franco, Bernhard Bermeitinger, Tiago Cunha, Brian Davis, André Freitas and Siegfried Handschuh

Humans naturally organise and classify the world into sets and categories. These categories expressed in natural language are present in all data artefacts from structured to unstructured data and play a fundamental role as tags, dataset predicates or ontology attributes. A better understanding of the category syntactic structure and how to match them semantically is a fundamental problem in the computational linguistics domain. Despite the high popularity of entity search, entity categories have not been receiving equivalent attention. This paper aims to present the task of semantic search of entity categories by defining, developing and making publicly available a multilingual test collection comprehending English, Portuguese and German. The test collections were designed to meet the demands of the entity search community in providing more representative and semantically complex query sets. In addition, we also provide comparative baselines and a brief analysis of the results.

Towards the Inference of Semantic Relations in Complex Nominals: a Pilot Study

Melania Cabezas-García and Pilar León-Araúz

Complex nominals (CNs) (e.g. wind turbine) are very common in English specialized texts (Nakov, 2013). However, all too frequently they show similar external forms but encode different semantic relations because of noun packing. This paper describes the use of paraphrases that convey the conceptual content of English two-term CNs (Nakov and Hearst, 2006) in the domain of environmental science. The semantic analysis of CNs was complemented by the use of knowledge patterns (KPs), which are lexico-syntactic patterns that usually convey semantic relations in real texts (Meyer, 2001; Marshman, 2006). Furthermore, the constituents of CNs were semantically annotated with conceptual categories (e.g. beach [LANDFORM] erosion [PROCESS]) with a view to disambiguating the semantic relation between the constituents of the CN and developing a procedure to infer the semantic relations in these multi-word terms. The results showed that the combination of KPs and paraphrases is a helpful approach to the semantics of CNs. Accordingly, the conceptual annotation of the constituents of CNs revealed similar patterns in the formation of these complex terms, which can lead to the inference of concealed semantic relations.

Generation of a Spanish Artificial Collocation Error Corpus

Sara Rodríguez-Fernández, Roberto Carlini and Leo Wanner

Collocations such as heavy rain or make [a] decision are combinations of two elements where one (the base) is freely chosen, while the choice of the other (collocate) is restricted by the base. Research has consistently shown that collocations present difficulties even to the most advanced language learners, so that computational tools aimed at supporting them in the process of language learning can be of great value. However, in contrast to grammatical error detection and correction, collocation error marking and correction has not yet received the attention it deserves. This is unsurprising, considering the lack of existing collocation resources, in particular those that capture the different types of collocation errors, and the high cost of a manual creation of such resources. In this paper, we present an algorithm for the automatic generation of an artificial collocation error corpus of American English learners of Spanish that includes 17 different types of collocation errors and that can be used for automatic detection and classification of collocation errors in the writings of Spanish language learners.

Improving a Neural-based Tagger for Multiword Expressions Identification

Dušan Variš and Natalia Klyueva

In this paper, we present a set of improvements introduced to MUMULS, a tagger for the automatic detection of verbal multiword expressions. Our tagger participated in the PARSEME shared task and it was the only one based on neural networks. We show that character-level embeddings can improve the performance, mainly by reducing the out-of-vocabulary rate. Furthermore, replacing the softmax layer in the decoder by a conditional random field classifier brings additional improvements. Finally, we compare different context-aware feature representations of input tokens using various

encoder architectures. The experiments on Czech show that the combination of character-level embeddings using a convolutional network, self-attentive encoding layer over the word representations and an output conditional random field classifier yields the best empirical results.

Designing a Russian Idiom-Annotated Corpus

Katsiaryna Aharodnik, Anna Feldman and Jing Peng

This paper describes the development of an idiom-annotated corpus of Russian. The corpus is compiled from freely available resources online and contains texts of different genres. The idiom extraction, annotation procedure, and a pilot experiment using the new corpus are outlined in the paper. Considering the scarcity of publicly available Russian annotated corpora, the corpus is a much-needed resource that can be utilized for literary, linguistic studies, pedagogy as well as for various Natural Language Processing tasks.

Session I-O2: Industry Track - Human computation in industry

10th May 2018, 14:50

Chair person: **Kyle Gorman** Oral Session

Academic-Industrial Perspective on the Development and Deployment of a Moderation System for a Newspaper Website

Dietmar Schabus and Marcin Skowron

This paper describes an approach and our experiences from the development, deployment and usability testing of a Natural Language Processing (NLP) and Information Retrieval system that supports the moderation of user comments on a large newspaper website. We highlight some of the differences between industry-oriented and academic research settings and their influence on the decisions made in the data collection and annotation processes, selection of document representation and machine learning methods. We report on classification results, where the problems to solve and the data to work with come from a commercial enterprise. In this context typical for NLP research, we discuss relevant industrial aspects. We believe that the challenges faced as well as the solutions proposed for addressing them can provide insights to others working in a similar setting.

Community-Driven Crowdsourcing: Data Collection with Local Developers

Christina Funk, Michael Tseng, Ravindran Rajakumar and Linne Ha

We tested the viability of partnering with local developers to create custom annotation applications and to recruit and motivate crowd contributors from their communities to perform an annotation task consisting of the assignment of toxicity ratings to Wikipedia comments. We discuss the background of the project, the design of the community-driven approach, the developers' execution of their applications and crowdsourcing programs, and the quantity, quality, and cost of judgments, in comparison with previous approaches. The community-driven approach resulted in local developers successfully creating four unique tools and collecting labeled data of sufficiently high quantity and quality. The creative approaches to the rating task presentation and crowdsourcing program design drew upon developers' local knowledge of their own social networks, who also reported interest in the underlying problem that the data collection addresses. We consider the lessons that may be drawn from this project for implementing future iterations of the community-driven approach.

Building Open Javanese and Sundanese Corpora for Multilingual Text-to-Speech

Jaka Aris Eko Wibawa, Supheakmungkol Sarin, Chenfang Li, Knot Pipatsrisawat, Keshan Sodimana, Oddur Kjartansson, Alexander Gutkin, Martin Jansche and Linne Ha

We present multi-speaker text-to-speech corpora for Javanese and Sundanese, the second and third largest languages of Indonesia spoken by well over a hundred million people. The key objectives were to collect high-quality data in an affordable way and to share the data publicly with the speech community. To achieve this, we collaborated with two local universities in Java and streamlined our recording and crowdsourcing processes to produce corpora consisting of 5,800 (Javanese) and 4,200 (Sundanese) mixed-gender recordings. We used these corpora to build several configurations of multi-speaker neural network-based text-to-speech systems for Javanese and Sundanese. Subjective evaluations performed on these configurations demonstrate that multilingual configurations for which Javanese and Sundanese are trained jointly with a larger corpus of Standard Indonesian significantly outperform the systems constructed from a single language. We hope that sharing these corpora publicly and presenting our multilingual approach to text-to-speech will help the community to scale up text-to-speech technologies to other lesser resourced languages of Indonesia.

Session O21 - Discourse & Argumentation

10th May 2018, 14:50

Chair person: **Costanza Navarretta** Oral Session

An Integrated Representation of Linguistic and Social Functions of Code-Switching

Silvana Hartmann, Monojit Choudhury and Kalika Bali

We present an integrated representation of code-switching (CS) functions, i.e., a representation that includes various CS phenomena (intra-/inter-sentential) and modalities (written/spoken), and aims to derive CS functions from local and global properties of the code-switched discourse. By applying it to several English/Hindi CS datasets, we show that our model contributes i) to the standardization and re-use of CS data collections by creating a resource footprint, and ii) to the study of CS functions by creating a systematic description and hierarchy of reported functions together with the (local and social) properties that may affect them. At the same time, the model provides a flexible framework to add emerging functions, supporting theoretical studies as well as the automatic detection of CS functions.

A Corpus of eRulemaking User Comments for Measuring Evaluability of Arguments

Joonsuk Park and Claire Cardie

eRulemaking is a means for government agencies to directly reach citizens to solicit their opinions and experiences regarding newly proposed rules. The effort, however, is partly hampered by citizens' comments that lack reasoning and evidence, which are largely ignored since government agencies are unable to evaluate the validity and strength. We present Cornell eRulemaking Corpus – CDCP, an argument mining corpus annotated with argumentative structure information capturing the evaluability of arguments. The corpus consists of 731 user comments on Consumer Debt Collection Practices (CDCP) rule by the Consumer Financial Protection Bureau (CFPB); the resulting dataset contains 4931 elementary unit and 1221 support relation annotations. It is a resource for building argument mining systems that can not only extract arguments from unstructured text, but also identify what additional information is necessary for readers to understand and evaluate a given argument. Immediate applications include providing real-time feedback to commenters, specifying which types of support for which propositions can be added to construct better-formed arguments.

A Multi-layer Annotated Corpus of Argumentative Text: From Argument Schemes to Discourse Relations

Elena Musi, Manfred Stede, Leonard Kriese, Smaranda Muresan and Andrea Rocci

We present a multi-layer annotated corpus of 112 argumentative microtexts encompassing not only argument structure and discourse relations (Stede et al., 2016), but also argument schemes — the inferential relations linking premises to claims. We propose a set of guidelines for the annotation of argument schemes both for support and attack relations, and a new user-friendly annotation tool. The multi-layer annotated corpus allows us to conduct an initial study of dependencies between discourse relations (according to Rhetorical Structure Theory (Mann and Thompson, 1988)) and argument schemes. Our main contribution is that of offering the first resource for the combined study of (argumentative) discourse relations and inferential moves.

Discourse Coherence Through the Lens of an Annotated Text Corpus: A Case Study

Eva Hajicova and Jiří Mírovský

A corpus-based study of local coherence as established by anaphoric links between the elements in the thematic (Topic) and the rhematic (Focus) parts of sentences in different genres of discourse. The study uses the Czech data present in the Prague Dependency Treebank and annotated for surface and underlying syntactic relations, the contextual boundness of tree nodes (from which the bi-partition of the sentence into Topic and Focus can be derived) and the coreference and bridging relations. Among the four possible types of the relations between anaphoric links and the Topic–Focus bipartition of the sentence, the most frequently occurring type is a link between the Topic of the sentence to the Focus of the immediately preceding sentence. In case there is an anaphoric link leading from the Focus of one sentence to the Topic or Focus of the immediately preceding sentence, this link frequently leads from a contextually bound element of the Focus, which supports the assumption that it is convenient to distinguish between "overall" Topic and Focus and the local Topic and Focus and/or the anaphoric relation is of the type of bridging and the relationship is often interpreted as a contrast. As for the relationship between the relations of the Topic-to-Topic type, due to the word order typological difference for Czech and English, these relations in Czech are not at all related to the syntactic function of subject.

Automatic Prediction of Discourse Connectives

Eric Malmi, Daniele Pighin, Sebastian Krause and Mikhail Kozhevnikov

Accurate prediction of suitable discourse connectives (however, furthermore, etc.) is a key component of any system aimed at

building coherent and fluent discourses from shorter sentences and passages. As an example, a dialog system might assemble a long and informative answer by sampling passages extracted from different documents retrieved from the Web. We formulate the task of discourse connective prediction and release a dataset of 2.9M sentence pairs separated by discourse connectives for this task. Then, we evaluate the hardness of the task for human raters, apply a recently proposed decomposable attention (DA) model to this task and observe that the automatic predictor has a higher F1 than human raters (32 vs. 30). Nevertheless, under specific conditions the raters still outperform the DA model, suggesting that there is headroom for future improvements.

Session O22 - Less-Resourced & Ancient Languages

10th May 2018, 14:50

Chair person: **Mark Liberman** Oral Session

Handling Rare Word Problem using Synthetic Training Data for Sinhala and Tamil Neural Machine Translation

Pasindu Tennage, Prabath Sandaruwan, Malith Thilakarathne, Achini Herath and Surangika Ranathunga

Lack of parallel training data influences the rare word problem in Neural Machine Translation (NMT) systems, particularly for under-resourced languages. Using synthetic parallel training data (data augmentation) is a promising approach to handle the rare word problem. Previously proposed methods for data augmentation do not consider language syntax when generating synthetic training data. This leads to generation of sentences that lower the overall quality of parallel training data. In this paper, we discuss the suitability of using Parts of Speech (POS) tagging and morphological analysis as syntactic features to prune the generated synthetic sentence pairs that do not adhere to language syntax. Our models show an overall 2.16 and 5.00 BLEU score gains over our benchmark Sinhala to Tamil and Tamil to Sinhala translation systems, respectively. Although we focus on Sinhala and Tamil NMT for the domain of official government documents, we believe that these synthetic data pruning techniques can be generalized to any language pair.

BDPROTO: A Database of Phonological Inventories from Ancient and Reconstructed Languages

Egidio Marsico, Sebastien Flavier, Annemarie Verkerk and Steven Moran

Here we present BDPROTO, a database comprised of phonological inventory data from 137 ancient and reconstructed languages. These data were extracted from historical linguistic reconstructions and brought together into a single unified, normalized, accessible, and Unicode-compliant language resource. This dataset is publicly available and we aim to engage language scientists doing research on language change and language evolution. We provide a short case study to highlight BDPROTO's research viability; using phylogenetic comparative methods and high-resolution language family trees, we investigate whether consonantal and vocalic systems differ in their rates of change over the last 10,000 years.

Creating a Translation Matrix of the Bible's Names Across 591 Languages

Winston Wu, Nidhi Vyas and David Yarowsky

For many of the world's languages, the Bible is the only significant bilingual, or even monolingual, text, making it a unique training resource for tasks such as translation, named entity analysis, and transliteration. Given the Bible's small size, however, the output of standard word alignment tools can be extremely noisy, making downstream tasks difficult. In this work, we develop and release a novel resource of 1129 aligned Bible person and place names across 591 languages, which was constructed and improved using several approaches including weighted edit distance, machine-translation-based transliteration models, and affixal induction and transformation models. Our models outperform a widely used word aligner on 97% of test words, showing the particular efficacy of our approach on the impactful task of broadly multilingual named-entity alignment and translation across a remarkably large number of world languages. We further illustrate the utility of our translation matrix for the multilingual learning of name-related affixes and their semantics as well as transliteration of named entities.

Building a Word Segmenter for Sanskrit Overnight

Vikas Reddy, Amrith Krishna, Vishnu Sharma, Prateek Gupta, Vineeth M R and Pawan Goyal

There is an abundance of digitised texts available in Sanskrit. However, the word segmentation task in such texts are challenging due to the issue of 'Sandhi'. In Sandhi, words in a sentence often fuse together to form a single chunk of text, where the word delimiter vanishes and sounds at the word boundaries undergo transformations, which is also reflected in the written text. Here, we propose an approach that uses a deep sequence to sequence (seq2seq) model that takes only the sandhied string as the input and predicts the unsandhied string. The state of the art models are linguistically involved and have external dependencies for the lexical and morphological analysis of the input. Our model can

be trained "overnight" and be used for production. In spite of the knowledge-lean approach, our system performs better than the current state of the art by gaining a percentage increase of 16.79 % than the current state of the art.

Simple Semantic Annotation and Situation Frames: Two Approaches to Basic Text Understanding in LORELEI

Kira Griffitt, Jennifer Tracey, Ann Bies and Stephanie Strassel

We present two types of semantic annotation developed for the DARPA Low Resource Languages for Emerging Incidents (LORELEI) program: Simple Semantic Annotation (SSA) and Situation Frames (SF). Both of these annotation approaches are concerned with labeling basic semantic information relevant to humanitarian aid and disaster relief (HADR) scenarios, with SSA serving as a more general resource and SF more directly supporting the evaluation of LORELEI technology. Mapping between information in different annotation tasks is an area of ongoing research for both system developers and data providers. We discuss the similarities and differences between the two types of LORELEI semantic annotation, along with ways in which the general semantic information captured in SSA can be leveraged in order to recognize HADR-oriented information captured by SF. To date we have produced annotations for nineteen LORELEI languages; by the program's end both SF and SSA will be available for over two dozen typologically diverse languages. Initially data is provided to LORELEI performers and to participants in NIST's Low Resource Human Language Technologies (LoReHLT) evaluation series. After their use in LORELEI and LoReHLT evaluations the data sets will be published in the LDC catalog.

Session O23 - Semantics & Evaluation

10th May 2018, 14:50

Chair person: **Gerard de Melo** Oral Session

Abstract Meaning Representation of Constructions: The More We Include, the Better the Representation

Claire Bonial, Bianca Badarau, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Tim O'Gorman, Martha Palmer and Nathan Schneider

We describe the expansion of the Abstract Meaning Representation (AMR) project to provide coverage for the annotation of certain types of constructions. Past AMR annotations generally followed a practice of assigning the semantic roles associated with an individual lexical item, as opposed to a flexible pattern or template of multiple lexical items, which characterizes constructions such as 'The X-er, The Y-er' (exemplified in the title). Furthermore, a goal of AMR is to provide consistent semantic representation despite language-specific syntactic idiosyncracies. Thus, representing the meanings associated with fully syntactic patterns required a novel annotation approach. As one strategy in our approach, we expanded the AMR lexicon of predicate senses, or semantic 'rolesets,' to include entries for a growing set of constructions. Despite the challenging practical and theoretical questions encountered, the additions and updates to AMR annotation described here ensure more comprehensive semantic representations capturing both lexical and constructional meaning.

Evaluating Scoped Meaning Representations

Rik Van Noord, Lasha Abzianidze, Hessel Haagsma and Johan Bos

Semantic parsing offers many opportunities to improve natural language understanding. We present a semantically annotated parallel corpus for English, German, Italian, and Dutch where sentences are aligned with scoped meaning representations in order to capture the semantics of negation, modals, quantification, and presupposition triggers. The semantic formalism is based on Discourse Representation Theory, but concepts are represented by WordNet synsets and thematic roles by VerbNet relations. Translating scoped meaning representations to sets of clauses enables us to compare them for the purpose of semantic parser evaluation and checking translations. This is done by computing precision and recall on matching clauses, in a similar way as is done for Abstract Meaning Representations. We show that our matching tool for evaluating scoped meaning representations is both accurate and efficient. Applying this matching tool to three baseline semantic parsers yields F-scores between 43% and 54%. A pilot study is performed to automatically find changes in meaning by comparing meaning representations of translations. This comparison turns out to be an additional way of (i) finding annotation mistakes and (ii) finding instances where our semantic analysis needs to be improved.

Huge Automatically Extracted Training-Sets for Multilingual Word SenseDisambiguation

Tommaso Pasini, Francesco Elia and Roberto Navigli

We release to the community six large-scale sense-annotated datasets in multiple language to pave the way for supervised multilingual Word Sense Disambiguation. Our datasets cover all the nouns in the English WordNet and their translations in other languages for a total of millions of sense-tagged sentences. Experiments prove that these corpora can be effectively used as

training sets for supervised WSD systems, surpassing the state of the art for low-resourced languages and providing competitive results for English, where manually annotated training sets are available. The data is available at trainomatic.org.

SentEval: An Evaluation Toolkit for Universal Sentence Representations

Alexis Conneau and Douwe Kiela

We introduce SentEval, a toolkit for evaluating the quality of universal sentence representations. SentEval encompasses a variety of tasks, including binary and multi-class classification, natural language inference and sentence similarity. The set of tasks was selected based on what appears to be the community consensus regarding the appropriate evaluations for universal sentence representations. The toolkit comes with scripts to download and preprocess datasets, and an easy interface to evaluate sentence encoders. The aim is to provide a fairer, less cumbersome and more centralized way for evaluating sentence representations.

A Survey on Automatically-Constructed WordNets and their Evaluation: Lexical and Word Embedding-based Approaches

Steven Neale

WordNets - lexical databases in which groups of synonyms are arranged according to the semantic relationships between them - are crucial resources in semantically-focused natural language processing tasks, but are extremely costly and labour intensive to produce. In languages besides English, this has led to growing interest in constructing and extending WordNets automatically, as an alternative to producing them from scratch. This paper describes various approaches to constructing WordNets automatically - by leveraging traditional lexical resources and newer trends such as word embeddings - and also offers a discussion of the issues affecting the evaluation of automatically constructed WordNets.

Session O24 - Multimodal & Written Corpora

10th May 2018, 14:50

Chair person: **Erhard Hinrichs** Oral Session

Linguistically-driven Framework for Computationally Efficient and Scalable Sign Recognition

Dimitri Metaxas, Mark Dilsizian and Carol Neidle

We introduce a new general framework for sign recognition from monocular video using limited quantities of annotated data. The novelty of the hybrid framework we describe here is that we exploit state-of-the art learning methods while also incorporating features based on what we know about the linguistic composition of lexical signs. In particular, we analyze hand shape, orientation, location, and motion trajectories, and then use CRFs to combine this linguistically significant information for purposes of sign recognition. Our robust modeling and recognition of these sub-components of sign production allow an efficient parameterization of the sign recognition problem as compared with purely data-driven methods. This parameterization enables a scalable and extendable time-series learning approach that advances the state of the art in sign recognition, as shown by the results reported here for recognition of isolated, citation-form, lexical signs from American Sign Language (ASL).

CONDUCT: An Expressive Conducting Gesture Dataset for Sound Control

Lei Chen, Sylvie Gibet and Camille Marteau

Recent research in music-gesture relationship has paid more attention on the sound variations and its corresponding gesture expressiveness. In this study we are interested by gestures performed by orchestral conductors, with a focus on the expressive gestures made by the non dominant hand. We make the assumption that these gestures convey some meaning shared by most of conductors, and that they implicitly correspond to sound effects which can be encoded in musical scores. Following this hypothesis, we defined a collection of gestures for musical direction. These gestures are designed to correspond to well known functional effect on sounds, and they can be modulated to vary this effect by simply modifying one of their structural component (hand movement or hand shape). This paper presents the design of the gesture and sound sets and the protocol that has led to the database construction. The relevant musical excerpts and the related expressive gestures have been first defined by one expert musician. The gestures were then recorded through motion capture by two non experts who performed them along with recorded music. This database will serve as a basis for training gesture recognition system for live sound control and modulation.

Neural Caption Generation for News Images

Vishwash Batra, Yulan He and George Vogiatzis

Automatic caption generation of images has gained significant interest. It gives rise to a lot of interesting image-related applications. For example, it could help in image/video retrieval and management of the vast amount of multimedia data available on the Internet. It could also help in the development of tools that can aid visually impaired individuals in accessing multimedia content. In this paper, we particularly focus on news images

and propose a methodology for automatically generating captions for news paper articles consisting of a text paragraph and an image. We propose several deep neural network architectures built upon Recurrent Neural Networks. Results on a BBC News dataset show that our proposed approach outperforms a traditional method based on Latent Dirichlet Allocation using both automatic evaluation based on BLEU scores and human evaluation.

MPST: A Corpus of Movie Plot Synopses with Tags

Sudipta Kar, Suraj Maharjan, Adrian Pastor López Monroy and Thamar Solorio

Social tagging of movies reveals a wide range of heterogeneous information about movies, like the genre, plot structure, soundtracks, metadata, visual and emotional experiences. Such information can be valuable in building automatic systems to create tags for movies. Automatic tagging systems can help recommendation engines to improve the retrieval of similar movies as well as help viewers to know what to expect from a movie in advance. In this paper, we set out to the task of collecting a corpus of movie plot synopses and tags. We describe a methodology that enabled us to build a fine-grained set of around 70 tags exposing heterogeneous characteristics of movie plots and the multi-label associations of these tags with some 14K movie plot synopses. We investigate how these tags correlate with movies and the flow of emotions throughout different types of movies. Finally, we use this corpus to explore the feasibility of inferring tags from plot synopses. We expect the corpus will be useful in other tasks where analysis of narratives is relevant.

OpenSubtitles2018: Statistical Rescoring of Sentence Alignments in Large, Noisy Parallel Corpora

Pierre Lison, Jörg Tiedemann and Milen Kouylekov

Movie and TV subtitles are a highly valuable resource for the compilation of parallel corpora thanks to their availability in large numbers and across many languages. However, the quality of the resulting sentence alignments is often lower than for other parallel corpora. This paper presents a new major release of the OpenSubtitles collection of parallel corpora, which is extracted from a total of 3.7 million subtitles spread over 60 languages. In addition to a substantial increase in the corpus size (about 30% compared to the previous version), this new release associates explicit quality scores to each sentence alignment. These scores are determined by a feedforward neural network based on simple language-independent features and estimated on a sample of aligned sentence pairs. Evaluation results show that the model is able predict lexical translation probabilities with a root mean square error of 0.07 (coefficient of determination R2 = 0.47). Based on the scores produced by this regression model, the parallel corpora can be filtered to prune out low-quality alignments.

Session P32 - Document Classification, Text Categorisation (1)

10th May 2018, 14:50

Chair person: **Monserrat Marimon** Poster Session

DeepTC – An Extension of DKPro Text Classification for Fostering Reproducibility of Deep Learning Experiments

Tobias Horsmann and Torsten Zesch

We present a deep learning extension for the multi-purpose text classification framework DKPro Text Classification (DKPro TC). DKPro TC is a flexible framework for creating easily shareable and reproducible end-to-end NLP experiments involving machine learning. We provide an overview of the current state of DKPro TC, which does not allow integration of deep learning, and discuss the necessary conceptual extensions. These extensions are based on an analysis of common deep learning setups found in the literature to support all common text classification setups, i.e. single outcome, multi outcome, and sequence classification problems. Additionally to providing an end-to-end shareable environment for deep learning experiments, we provide convenience features that take care of repetitive steps, such as pre-processing, data vectorization and pruning of embeddings. By moving a large part of this boilerplate code into DKPro TC, the actual deep learning framework code improves in readability and lowers the amount of redundant source code considerably. As proof-of-concept, we integrate Keras, DyNet, and DeepLearning4J.

Improving Hate Speech Detection with Deep Learning Ensembles

Steven Zimmerman, Udo Kruschwitz and Chris Fox

Hate speech has become a major issue that is currently a hot topic in the domain of social media. Simultaneously, current proposed methods to address the issue raise concerns about censorship. Broadly speaking, our research focus is the area human rights, including the development of new methods to identify and better address discrimination while protecting freedom of expression. As neural network approaches are becoming state of the art for text classification problems, an ensemble method is adapted for usage with neural networks and is presented to better classify hate speech. Our method utilizes a publicly available embedding model, which is tested against a hate speech corpus from Twitter. To confirm robustness of our results, we additionally test against

a popular sentiment dataset. Given our goal, we are pleased that our method has a nearly 5 point improvement in F-measure when compared to original work on a publicly available hate speech evaluation dataset. We also note difficulties encountered with reproducibility of deep learning methods and comparison of findings from other work. Based on our experience, more details are needed in published work reliant on deep learning methods, with additional evaluation information a consideration too. This information is provided to foster discussion within the research community for future work.

Distributional Term Set Expansion

Amaru Cuba Gyllensten and Magnus Sahlgren

This paper is a short empirical study of the performance of centrality and classification based iterative term set expansion methods for distributional semantic models. Iterative term set expansion is an interactive process using distributional semantics models where a user labels terms as belonging to some sought after term set, and a system uses this labeling to supply the user with new, candidate, terms to label, trying to maximize the number of positive examples found. While centrality based methods have a long history in term set expansion, we compare them to classification methods based on the the Simple Margin method, an Active Learning approach to classification using Support Vector Machines. Examining the performance of various centrality and classification based methods for a variety of distributional models over five different term sets, we can show that active learning based methods consistently outperform centrality based methods.

Can Domain Adaptation be Handled as Analogies?

Núria Bel and Joel Pocostales

Aspect identification in user generated texts by supervised text classification might suffer degradation in performance when changing to other domains than the one used for training. For referring to aspects such as quality, price or customer services the vocabulary might differ and affect performance. In this paper, we present an experiment to validate a method to handle domain shifts when there is no available labeled data to retrain. The system is based on the offset method as used for solving word analogy problems in vector semantic models such as word embedding. Despite of the fact that the offset method indeed found relevant analogues in the new domain for the classifier initial selected features, the classifiers did not deliver the expected results. The analysis showed that a number of words were found as analogues for many different initial features. This phenomenon was already described in the literature as 'default words' or 'hubs'. However, our data showed that it cannot be explained in terms of word frequency or distance to the question word, as suggested.

Author Profiling from Facebook Corpora

Fernando Hsieh, Rafael Dias and Ivandré Paraboni

Author profiling - the computational task of prediction author's demographics from text - has been a popular research topic in the NLP field, and also the focus of a number of prominent shared tasks. Author profiling is a problem of growing importance, with applications in forensics, security, sales and many others. In recent years, text available from social networks has become a primary source for computational models of author profiling, but existing studies are still largely focused on age and gender prediction, and are in many cases limited to the use of English text. Other languages, and other author profiling tasks, remain somewhat less popular. As a means to further this issue, in this work we present initial results of a number of author profiling tasks from a Facebook corpus in the Brazilian Portuguese language. As in previous studies, our own work will focus on both standard gender and age prediction tasks but, in addition to these, we will also address two less usual author profiling tasks, namely, predicting an author's degree of religiosity and IT background status. The tasks are modelled by making use of different knowledge sources, and results of alternative approaches are discussed.

Semantic Relatedness of Wikipedia Concepts – Benchmark Data and a Working Solution

Liat Ein Dor, Alon Halfon, Yoav Kantor, Ran Levy, Yosi Mass, Ruty Rinott, Eyal Shnarch and Noam Slonim

Wikipedia is a very popular source of encyclopedic knowledge which provides highly reliable articles in a variety of domains. This richness and popularity created a strong motivation among NLP researchers to develop relatedness measures between Wikipedia concepts. In this paper, we introduce WORD (Wikipedia Oriented Relatedness Dataset), a new type of concept relatedness dataset, composed of 19,276 pairs of Wikipedia concepts. This is the first human annotated dataset of Wikipedia concepts, whose purpose is twofold. On the one hand, it can serve as a benchmark for evaluating concept-relatedness methods. On the other hand, it can be used as supervised data for developing new models for concept relatedness prediction. Among the advantages of this dataset compared to its term-relatedness counterparts, are its built-in disambiguation solution, and its richness with meaningful multiword terms. Based on this benchmark we develop a new tool, named WORT (Wikipedia Oriented Relatedness Tool), for measuring the level of relatedness between pairs of concepts. We show that the relatedness predictions ofWORT outperform state of the art methods.

Experiments with Convolutional Neural Networks for Multi-Label Authorship Attribution

Dainis Boumber, Yifan Zhang and Arjun Mukherjee

We explore the use of Convolutional Neural Networks (CNNs) for multi-label Authorship Attribution (AA) problems and propose a CNN specifically designed for such tasks. By averaging the author probability distributions at sentence level for the longer documents and treating smaller documents as sentences, our multi-label design adapts to single-label datasets and various document sizes, retaining the capabilities of a traditional CNN. As a part of this work, we also create and make available to the public a multi-label Authorship Attribution dataset (MLPA-400), consisting of 400 scientific publications by 20 authors from the field of Machine Learning. Proposed Multi-label CNN is evaluated against a large number of algorithms on MLPA-400 and PAN-2012, a traditional single-label AA benchmark dataset. Experimental results demonstrate that our method outperforms several state-of-the-art models on the proposed task.

Session P33 - Morphology (1)

10th May 2018, 14:50

Chair person: **Piotr Banski** Poster Session

A Fast and Accurate Vietnamese Word Segmenter

Dat Quoc Nguyen, Dai Quoc Nguyen, Thanh Vu, Mark Dras and Mark Johnson

We propose a novel approach to Vietnamese word segmentation. Our approach is based on the Single Classification Ripple Down Rules methodology (Compton and Jansen, 1990), where rules are stored in an exception structure and new rules are only added to correct segmentation errors given by existing rules. Experimental results on the benchmark Vietnamese treebank show that our approach outperforms previous state-of-the-art approaches JVnSegmenter, vnTokenizer, DongDu and UETsegmenter in terms of both accuracy and performance speed. Our code is open-source and available at: https://github.com/datquocnguyen/RDRsegmenter.

Finite-state morphological analysis for Gagauz

Francis Tyers, Sevilay Bayatli, Güllü Karanfil and Memduh Gokirmak

This paper describes a finite-state approach to morphological analysis and generation of Gagauz, a Turkic language spoken in the Republic of Moldova. Finite-state approaches are commonly used in morphological modelling, but one of the novelties of our approach is that we explicitly handle orthographic errors and variance, in addition to loan words. The resulting model has a reasonable coverage (above 90\%) over a range of freely-available corpora.

Albanian Part-of-Speech Tagging: Gold Standard and Evaluation

Besim Kabashi and Thomas Proisl

In this paper, we present a gold standard corpus for Albanian part-of-speech tagging and perform evaluation experiments with different statistical taggers. The corpus consists of more than 31,000 tokens and has been manually annotated with a medium-sized tagset that can adequately represent the syntagmatic aspects of the language. We provide mappings from the full tagset to both the original Google Universal Part-of-Speech Tags and the variant used in the Universal Dependencies project. We perform experiments with different taggers on the full tagset as well as on the coarser tagsets and achieve accuracies of up to 95.10%.

Morphology Injection for English-Malayalam Statistical Machine Translation

Sreelekha S and Pushpak Bhattacharyya

Statistical Machine Translation (SMT) approaches fails to handle the rich morphology when translating into morphologically rich languages. This is due to the data sparsity, which is the missing of the morphologically inflected forms of words from the parallel corpus. We investigated a method to generate these unseen morphological forms. In this paper, we analyze the morphological complexity of a morphologically rich Indian language Malayalam when translating from English. Being a highly agglutinative language, it is very difficult to generate the various morphological inflected forms for Malayalam. We study both the factor based models and the phrase based models and the problem of data sparseness. We propose a simple and effective solution based on enriching the parallel corpus with generated morphological forms. We verify this approach with various experiments on English-Malayalam SMT. We observes that the morphology injection method improves the quality of the translation. We have analyzed the experimental results both in terms of automatic and subjective evaluations.

The Morpho-syntactic Annotation of Animacy for a Dependency Parser

Mohammed Attia, Vitaly Nikolaev and Ali Elkahky

In this paper we present the annotation scheme and parser results of the animacy feature in Russian and Arabic, two morphologicallyrich languages, in the spirit of the universal dependency framework (McDonald et al., 2013; de Marneffe et al., 2014). We explain the animacy hierarchies in both languages and make the case for the existence of five animacy types. We train a morphological analyzer on the annotated

data and the results show a prediction f-measure for animacy of 95.39% for Russian and 92.71% for Arabic. We also use animacy along with other morphological tags as features to train a dependency parser, and the results show a slight improvement gained from animacy. We compare the impact of animacy on improving the dependency parser to other features found in nouns, namely, 'gender', 'number', and 'case'. To our knowledge this is the first contrastive study of the impact of morphological features on the accuracy of a transition parser. A portion of our data (1,000 sentences for Arabic and Russian each, along with other languages) annotated according to the scheme described in this paper is made publicly available (https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1983) as part of the CoNLL 2017 Shared Task on Multilingual Parsing (Zeman et al., 2017).

MADARi: A Web Interface for Joint Arabic Morphological Annotation and Spelling Correction

Ossama Obeid, Salam Khalifa, Nizar Habash, Houda Bouamor, Wajdi Zaghouani and Kemal Oflazer

In this paper, we introduce MADARi, a joint morphological annotation and spelling correction system for texts in Standard and Dialectal Arabic. The MADARi framework provides intuitive interfaces for annotating text and managing the annotation process of a large number of sizable documents. Morphological annotation includes indicating, for a word, in context, its baseword, clitics, part-of-speech, lemma, gloss, and dialect identification. MADARi has a suite of utilities to help with annotator productivity. For example, annotators are provided with pre-computed analyses to assist them in their task and reduce the amount of work needed to complete it. MADARi also allows annotators to query a morphological analyzer for a list of possible analyses in multiple dialects or look up previously submitted analyses. The MADARi management interface enables a lead annotator to easily manage and organize the whole annotation process remotely and concurrently. We describe the motivation, design and implementation of this interface; and we present details from a user study working with this system.

A Morphological Analyzer for St. Lawrence Island / Central Siberian Yupik

Emily Chen and Lane Schwartz

St. Lawrence Island / Central Siberian Yupik is an endangered language, indigenous to St. Lawrence Island in Alaska and the Chukotka Peninsula of Russia, that exhibits pervasive agglutinative and polysynthetic properties. This paper discusses an implementation of a finite-state morphological analyzer for Yupik that was developed in accordance with the grammatical standards and phenomena documented in Jacobson's 2001 reference grammar for Yupik. The analyzer was written in foma, an open source framework for constructing finite-state grammars of morphology. The approach presented here cyclically interweaves morphology and phonology to account for the language's intricate morphophonological system, an approach that may be applicable to typologically similar languages. The morphological analyzer has been designed to serve as foundational resource that will eventually underpin a suite of computational tools for Yupik to assist in the process of linguistic documentation and revitalization.

Universal Morphologies for the Caucasus region

Christian Chiarcos, Kathrin Donandt, Maxim Ionov, Monika Rind-Pawlowski, Hasmik Sargsian, Jesse Wichers Schreur, Frank Abromeit and Christian Fäth

The Caucasus region is famed for its rich and diverse arrays of languages and language families, often challenging European-centered views established in traditional linguistics. In this paper, we describe ongoing efforts to improve the coverage of Universal Morphologies for languages of the Caucasus region. The Universal Morphologies (UniMorph) are a recent community project aiming to complement the Universal Dependencies which focus on morphosyntax and syntax. We describe the development of UniMorph resources for Nakh-Daghestanian and Kartvelian languages as a well as for Classical Armenian, we discuss challenges that the complex morphology of these and related languages poses to the current design of UniMorph, and suggest possibilities to improve the applicability of UniMorph for languages of the Caucasus region in particular and for low resource languages in general. We also criticize the UniMorph TSV format for its limited expressiveness, and suggest to complement the existing UniMorph workflow with support for additional source formats on grounds of Linked Open Data technology.

Session P34 - Opinion Mining / Sentiment Analysis (2)

10th May 2018, 14:50

Chair person: **Patrick Paroubek** Poster Session

EMTC: Multilabel Corpus in Movie Domain for Emotion Analysis in Conversational Text

Phan Duc-Anh and Yuji Matsumoto

It is proved that in text-based communication such as sms, messengers applications, misinterpretation of partner's emotions

are pretty common. In order to tackle this problem, we propose a new multilabel corpus named Emotional Movie Transcript Corpus (EMTC). Unlike most of the existing emotion corpora that are collected from Twitters and use hashtags labels, our corpus includes conversations from movie with more than 2.1 millions utterances which are partly annotated by ourselves and independent annotators. To our intuition, conversations from movies are closer to real-life settings and emotionally richer. We believe that a corpus like EMTC will greatly benefit the development and evaluation of emotion analysis systems and improve their ability to express and interpret emotions in text-based communication.

Complex and Precise Movie and Book Annotations in French Language for Aspect Based Sentiment Analysis

Stefania Pecore and Jeanne Villaneau

Aspect Based Sentiment Analysis (ABSA) aims at collecting detailed opinion information according to products and their features, via the recognition of targets of the opinions in text. Though some annotated data have been produced in challenges as SemEval, resources are still scarce, especially for languages other than English. We are interested in enhancing today's mostly statistical text classification with the use of linguistics tools, in order to better define and analyze what has been written. The work presented in this paper focuses on two French datasets of movies and books online reviews. In reviews, text length is much higher compared to a tweet, giving us the opportunity to work on a challenging and linguistically interesting dataset. Moreover, movies and books are products that make classifying opinions into aspects quite complex. This article provides an analysis of the particularities of the two domains during the process of collecting and annotating data, a precise annotation scheme for each domain, examples and statistics issued from the annotation phase, and some perspectives on our future work.

Lingmotif-lex: a Wide-coverage, State-of-the-art Lexicon for Sentiment Analysis

Antonio Moreno-Ortiz and Chantal Pérez-Hernández

We present Lingmotif-lex, a new, wide-coverage, domain-neutral lexicon for sentiment analysis in English. We describe the creation process of this resource, its assumptions, format, and valence system. Unlike most sentiment lexicons currently available, Lingmotif-lex places strong emphasis on multi-word expressions, and has been manually curated to be as accurate, unambiguous, and comprehensive as possible. Also unlike existing available resources, \textit{Lingmotif-lex} comprises a comprehensive set of contextual valence shifters (CVS) that account for valence modification by context. Formal evaluation is provided by testing it on two publicly available sentiment analysis datasets, and comparing it with other English sentiment lexicons available, which we adapted to make this comparison as fair as possible. We show how Lingmotif-lex achieves significantly better performance than these lexicons across both datasets.

A Japanese Corpus for Analyzing Customer Loyalty Information

Yiou Wang and Takuji Tahara

Today customers voice attitudes, opinions and their experience about some brands, companies, products or services through center calls, web reviews or SNS and analyzing them is an important task. On the other hand, customer loyalty has long been a topic of high interest in both academia and industry. Therefore, it is attractive to consider exploiting customer loyalty information by analyzing the voice of customer. However, although many previous studies focused on analyzing attitudes, opinions, sentiments of the text data, no work has been conducted from the perspective of customer loyalty, which is reflected by a combination of customer attitudes and behavior. In this work, we present JCLIC, Japanese Customer Loyalty Information Corpus, which is a corpus for analyzing customer loyalty information. For each review we have annotated detailed customer loyalty information which contains: loyalty degree that reflects loyalty level of the customer, loyalty expression that expresses the customer loyalty, loyalty type that indicates the category to which loyalty expression belongs., reason expression that expresses why the customer have such loyalty degree, and reason type that indicates the category to which reason expression belongs.. We describe our annotation scheme and annotation process, present results of an agreement study and give some statistics about the corpus we have annotated.

FooTweets: A Bilingual Parallel Corpus of World Cup Tweets

Henny Sluyter-Gäthje, Pintu Lohar, Haithem Afli and Andy Way

The way information spreads through society has changed significantly over the past decade with the advent of online social networking. Twitter, one of the most widely used social networking websites, is known as the real-time, public microblogging network where news breaks first. Most users love it for its iconic 140-character limitation and unfiltered feed that show them news and opinions in the form of tweets. Tweets are usually multilingual in nature and of varying quality. However, machine translation (MT) of twitter data is a challenging task especially due to the following two reasons: (i) tweets are informal in nature (i.e., violates linguistic norms), and (ii) parallel resource for twitter data is scarcely available on the Internet. In this paper, we develop

FooTweets, a first parallel corpus of tweets for English–German language pair. We extract 4, 000 English tweets from the FIFA 2014 world cup and manually translate them into German with a special focus on the informal nature of the tweets. In addition to this, we also annotate sentiment scores between 0 and 1 to all the tweets depending upon the degree of sentiment associated with them. This data has recently been used to build sentiment translation engines and an extensive evaluation revealed that such a resource is very useful in machine translation of user generated content.

The SSIX Corpora: Three Gold Standard Corpora for Sentiment Analysis in English, Spanish and German Financial Microblogs

Thomas Gaillat, Manel Zarrouk, André Freitas and Brian Davis

This paper introduces the three SSIX corpora for sentiment analysis. These corpora address the need to provide annotated data for supervised learning methods. They focus on stock-market related messages extracted from two financial microblog platforms, i.e., StockTwits and Twitter. In total they include 2,886 messages with opinion targets. These messages are provided with polarity annotation set on a continuous scale by three or four experts in each language. The annotation information identifies the targets with a sentiment score. The annotation process includes manual annotation verified and consolidated by financial experts. The creation of the annotated corpora took into account principled sampling strategies as well as inter-annotator agreement before consolidation in order to maximize data quality.

Sarcasm Target Identification: Dataset and An Introductory Approach

Aditya Joshi, Pranav Goel, Pushpak Bhattacharyya and Mark Carman

Past work in computational sarcasm deals primarily with sarcasm detection. In this paper, we introduce a novel, related problem: sarcasm target identification (i.e., extracting the target of ridicule in a sarcastic sentence). As a benchmark, we introduce a new dataset for the task. This dataset is manually annotated for the sarcasm target in book snippets and tweets based on our formulation of the task. We then introduce an automatic approach for sarcasm target identification. It is based on a combination of two types of extractors: one based on rules, and another consisting of a statistical classifier. Our introductory approach establishes the viability of sarcasm target identification, and will serve as a baseline for future work.

Annotating Opinions and Opinion Targets in Student Course Feedback

Janaka Chathuranga, Shanika Ediriweera, Ravindu Hasantha, Pranidhith Munasinghe and Surangika Ranathunga

In this paper, we present a student course feedback corpus, a novel resource for opinion target extraction and sentiment analysis. The corpus is developed with the main aim of summarizing general feedback given by students on undergraduate-level courses. In this corpus, opinion targets, opinion expressions, and polarities of the opinion expressions towards the opinion targets are annotated. Opinion targets are basically the important key points in feedback that the students have shown their sentiment towards, such as “Lecture Slides”, and “Teaching Style”. The uniqueness of the corpus, annotation methodology, difficulties faced during annotating, and possible usages of the corpus are discussed in this paper.

Generating a Gold Standard for a Swedish Sentiment Lexicon

Jacobo Rouces, Nina Tahmasebi, Lars Borin and Stian Rødven Eide

There is an increasing demand for multilingual sentiment analysis, and most work on sentiment lexicons is still carried out based on English lexicons like WordNet. In addition, many of the non-English sentiment lexicons that do exist have been compiled by (machine) translation from English resources, thereby arguably obscuring possible language-specific characteristics of sentiment-loaded vocabulary. In this paper we describe the creation of a gold standard for the sentiment annotation of Swedish terms as a first step towards the creation of a full-fledged sentiment lexicon for Swedish – i.e., a lexicon containing information about \emph{prior} sentiment (also called polarity) values of lexical items (words or disambiguated word senses), along a scale negative–positive. We create a gold standard for sentiment annotation of Swedish terms, using the freely available SALDO lexicon and the Gigaword corpus. For this purpose, we employ a multi-stage approach combining corpus-based frequency sampling and two stages of human annotation: direct score annotation followed by Best-Worst Scaling. In addition to obtaining a gold standard, we analyze the data from our process and we draw conclusions about the optimal sentiment model.

Session P35 - Session Phonetic Databases, Phonology

10th May 2018, 14:50

Chair person: **Martine Adda-Decker** Poster Session

WordKit: a Python Package for Orthographic and Phonological Featurization

Stephan Tulkens, Dominiek Sandra and Walter Daelemans

The modeling of psycholinguistic phenomena, such as word reading, with machine learning techniques requires the featurization of word stimuli into appropriate orthographic and phonological representations. Critically, the choice of features impacts the performance of machine learning algorithms, and can have important ramifications for the conclusions drawn from a model. As such, featurizing words with a variety of feature sets, without having to resort to using different tools is beneficial development. In this work, we present wordkit, a python package which allows users to switch between feature sets and featurizers with a uniform API, allowing for rapid prototyping. To the best of our knowledge, this is the first package which integrates a variety of orthographic and phonological featurizers in a single package. The package is fully compatible with scikit-learn, and hence can be integrated into other pipelines. Furthermore, the package is modular and extensible, allowing for the integration of a large variety of feature sets and featurizers. The package and documentation can be found at github.com/stephantul/wordkit

Pronunciation Variants and ASR of Colloquial Speech: A Case Study on Czech

David Lukeš, Marie Kopřivová, Zuzana Komrsková and Petra Klimešová

A standard ASR system is built using three types of mutually related language resources: apart from speech recordings and orthographic transcripts, a pronunciation component maps tokens in the transcripts to their phonetic representations. Its implementation is either lexicon-based (whether by way of simple lookup or of a stochastic grapheme-to-phoneme converter trained on the source lexicon) or rule-based, or a hybrid thereof. Whichever approach ends up being taken (as determined primarily by the writing system of the language in question), little attention is usually paid to pronunciation variants stemming from connected speech processes, hypoarticulation, and other phenomena typical for colloquial speech, mostly because the resource is seldom directly empirically derived. This paper presents a case study on the automatic recognition of colloquial Czech, using a pronunciation dictionary extracted from the ORTOFON corpus of informal spontaneous Czech, which is manually phonetically transcribed. The performance of the dictionary is compared to a standard rule-based pronunciation component, as evaluated against a subset of the ORTOFON corpus (multiple speakers recorded on a single compact device) and the Vystadial telephone speech corpus, for which prior benchmarks are available.

Epitran: Precision G2P for Many Languages

David R. Mortensen, Siddharth Dalmia and Patrick Littell

Epitran is a massively multilingual, multiple back-end system for G2P (grapheme-to-phoneme) transduction which is distributed with support for 61 languages. It takes word tokens in the orthography of a language and outputs a phonemic representation in either IPA or X-SAMPA. The main system is written in Python and is publicly available as open source software. Its efficacy has been demonstrated in multiple research projects relating to language transfer, polyglot models, and speech. In a particular ASR task, Epitran was shown to improve the word error rate over Babel baselines for acoustic modeling.

Session P36 - Question Answering and Machine Reading

10th May 2018, 14:50

Chair person: **António Branco** Poster Session

A Multilingual Approach to Question Classification

Aikaterini-Lida Kalouli, Katharina Kaiser, Annette Hautli-Janisz, Georg A. Kaiser and Miriam Butt

In this paper we present the Konstanz Resource of Questions (KRoQ), the first dependency-parsed, parallel multilingual corpus of information-seeking and non-information-seeking questions. In creating the corpus, we employ a linguistically motivated rule-based system that uses linguistic cues from one language to help classify and annotate questions across other languages. Our current corpus includes German, French, Spanish and Koine Greek. Based on the linguistically motivated heuristics we identify, a two-step scoring mechanism assigns intra- and inter-language scores to each question. Based on these scores, each question is classified as being either information seeking or non-information seeking. An evaluation shows that this mechanism correctly classifies questions in 79% of the cases. We release our corpus as a basis for further work in the area of question classification. It can be utilized as training and testing data for machine-learning algorithms, as corpus-data for theoretical linguistic questions or as a resource for further rule-based approaches to question identification.

Dataset for the First Evaluation on Chinese Machine Reading Comprehension

Yiming Cui, Ting Liu, Zhipeng Chen, Wentao Ma, Shijin Wang and Guoping Hu

Machine Reading Comprehension (MRC) has become enormously popular recently and has attracted a lot of attention. However, existing reading comprehension datasets are mostly in English. To add diversity in reading comprehension datasets, in this paper we propose a new Chinese reading comprehension dataset for accelerating related research in the community. The proposed dataset contains two different types: cloze-style reading comprehension and user query reading comprehension, associated with large-scale training data as well as human-annotated validation and hidden test set. Along with this dataset, we also hosted the first Evaluation on Chinese Machine Reading Comprehension (CMRC-2017) and successfully attracted tens of participants, which suggest the potential impact of this dataset.

A Multi-Domain Framework for Textual Similarity. A Case Study on Question-to-Question and Question-Answering Similarity Tasks

Amir Hazem, Basma El Amel Boussaha and Nicolas Hernandez

Community Question Answering (CQA) websites have become a very popular and useful source of information, which helps users to find out answers to their corresponding questions. On one hand, if a user's question does not exist in the forum, a new post is created so that other users can contribute and provide answers or comments. On the other hand, if similar or related questions already exist in the forum, the system should be able to detect them and redirect the user towards the corresponding threads. This procedure of detecting similar questions is also known as question-to-question similarity task in the NLP research community. Once the correct posts have been detected, it is important to provide the correct answer since some posts can contain tens or hundreds of answers/comments which make the user's research more difficult. This procedure is also known as the question-answering similarity task. In this paper, we address both tasks and aim at providing the first framework on the evaluation of similar questions and question-answering detection on a multi-domain corpora. For that purpose, we use the community question answering forum Stack-Exchange to extract posts and pairs of questions and answers from multiple domains. We evaluate two baseline approaches over 19 domains and provide preliminary results on multiple annotated question-answering datasets to deal with question-answering similarity task.

WorldTree: A Corpus of Explanation Graphs for Elementary Science Questions supporting Multi-hop Inference

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein and Clayton Morrison

Developing methods of automated inference that are able to provide users with compelling human-readable justifications for why the answer to a question is correct is critical for domains such as science and medicine, where user trust and detecting costly errors are limiting factors to adoption. One of the central barriers to training question answering models on explainable inference tasks is the lack of gold explanations to serve as training data. In this paper we present a corpus of explanations for standardized science exams, a recent challenge task for question answering. We manually construct a corpus of detailed explanations for nearly all publicly available standardized elementary science question (approximately 1,680 3rd through 5th grade questions) and represent these as "explanation graphs" – sets of lexically overlapping sentences that describe how to arrive at the correct answer to a question through a combination of domain and world knowledge. We also provide an explanation-centered tablestore, a collection of semi-structured tables that contain the knowledge to construct these elementary science explanations. Together, these two knowledge resources map out a substantial portion of the knowledge required for answering and explaining elementary science exams, and provide both structured and free-text training data for the explainable inference task.

Analysis of Implicit Conditions in Database Search Dialogues

Shun-ya Fukunaga, Hitoshi Nishikawa, Takenobu Tokunaga, Hikaru Yokono and Tetsuro Takahashi

This paper reports an annotation to a corpus of database search dialogues on real estate, and the analysis on implicit information in the utterances for constructing database queries. Two annotators annotated 50 dialogues with a set of database field tags, resulting in a high inter-annotator agreement (Cohen's kappa=0.79), and the analysis revealed that 10% of the utterances included non-database-field information. We further investigated these utterances to find that more than 93% of them included useful information for figuring out search conditions, which we call the implicit condition. The contribution of this paper is to present the existence and importance of the implicit conditions in the database search dialogues and both qualitative and quantitative analysis of them. Our corpus can provide a fundamental language resource for constructing a dialogue system which can utilise the implicit conditions. The paper concluded with possible approaches to achieve our long-term goal, extracting the implicit

conditions in the database search dialogues and utilising them to construct queries.

An Information-Providing Closed-Domain Human-Agent Interaction Corpus

Jelte Van Waterschoot, Guillaume Dubuisson Duplessis, Lorenzo Gatti, Merijn Bruijnes and Dirk Heylen

The contribution of this paper is twofold: 1) we provide a public corpus for Human-Agent Interaction (where the agent is controlled by a Wizard of Oz) and 2) we show a study on verbal alignment in Human-Agent Interaction, to exemplify the corpus' use. In our recordings for the Human-Agent Interaction Alice-corpus (HAI Alice-corpus), participants talked to a wizarded agent, who provided them with information about the book Alice in Wonderland and its author. The wizard had immediate and almost full control over the agent's verbal and nonverbal behavior, as the wizard provided the agent's speech through his own voice and his facial expressions were directly copied onto the agent. The agent's hand gestures were controlled through a button interface. Data was collected to create a corpus with unexpected situations, such as misunderstandings, (accidental) false information, and interruptions. The HAI Alice-corpus consists of transcribed audio-video recordings of 15 conversations (more than 900 utterances) between users and the wizarded agent. As a use-case example, we measured the verbal alignment between the user and the agent. The paper contains information about the setup of the data collection, the unexpected situations and a description of our verbal alignment study.

Augmenting Image Question Answering Dataset by Exploiting Image Captions

Masashi Yokota and Hideki Nakayama

Image question answering (IQA) is one of the tasks that need rich resources, i.e. supervised data, to achieve optimal performance. However, because IQA is a challenging task that handles complex input and output information, the cost of naive manual annotation can be prohibitively expensive. On the other hand, it is thought to be relatively easy to obtain relevant pairs of an image and text in an unsupervised manner (e.g., crawling Web data). Based on this expectation, we propose a framework to augment training data for IQA by generating additional examples from unannotated pairs of an image and captions. The important constraint that a generated IQA example must satisfy is that its answer must be inferable from the corresponding image and question. To satisfy this, we first select a possible answer for a given image by randomly extracting an answer from corresponding captions. Then we generate the question from the triplets of the image, captions and fixed answer. In experiments, we test our method on the Visual Genome dataset varying the ratio of seed supervised data and demonstrate its effectiveness.

Semi-supervised Training Data Generation for Multilingual Question Answering

Kyungjae Lee, Kyoungho Yoon, Sunghyun Park and Seung-won Hwang

Recently, various datasets for question answering (QA) research have been released, such as SQuAD, Marco, WikiQA, MCTest, and SearchQA. However, such existing training resources for these task mostly support only English. In contrast, we study semi-automated creation of the Korean Question Answering Dataset (K-QuAD), by using automatically translated SQuAD and a QA system bootstrapped on a small QA pair set. As a naive approach for other language, using only machine-translated SQuAD shows limited performance due to translation errors. We study why such approach fails and motivate needs to build seed resources to enable leveraging such resources. Specifically, we annotate seed QA pairs of small size (4K) for Korean language, and design how such seed can be combined with translated English resources. These approach, by combining two resources, leads to 71.50 F1 on Korean QA (comparable to 77.3 F1 on SQuAD).

PhotoshopQuiA: A Corpus of Non-Factoid Questions and Answers for Why-Question Answering

Andrei Dulceanu, Thang Le Dinh, Walter Chang, Trung Bui, Doo Soon Kim, Manh Chien Vu and Seokhwan Kim

Recent years have witnessed a high interest in non-factoid question answering using Community Question Answering (CQA) web sites. Despite ongoing research using state-of-the-art methods, there is a scarcity of available datasets for this task. Why-questions, which play an important role in open-domain and domain-specific applications, are difficult to answer automatically since the answers need to be constructed based on different information extracted from multiple knowledge sources. We introduce the PhotoshopQuiA dataset, a new publicly available set of 2,854 why-question and answer(s) (WhyQ, A) pairs related to Adobe Photoshop usage collected from five CQA web sites. We chose Adobe Photoshop because it is a popular and well-known product, with a lively, knowledgeable and sizeable community. To the best of our knowledge, this is the first English dataset for Why-QA that focuses on a product, as opposed to previous open-domain datasets. The corpus is stored in JSON format and contains detailed data about questions and questioners as well as

answers and answerers. The dataset can be used to build Why-QA systems, to evaluate current approaches for answering why-questions, and to develop new models for future QA systems research.

BioRead: A New Dataset for Biomedical Reading Comprehension

Dimitris Pappas, Ion Androutsopoulos and Haris Papageorgiou

We present BioRead, a new publicly available cloze-style biomedical machine reading comprehension (MRC) dataset with approximately 16.4 million passage-question instances. BioRead was constructed in the same way as the widely used Children's Book Test and its extension BookTest, but using biomedical journal articles and employing MetaMap to identify UMLS concepts. BioRead is one of the largest MRC datasets, and currently the largest one in the biomedical domain. We also provide a subset of BioRead, BioReadLite, for research groups with fewer computational resources. We re-implemented and tested on BioReadLite two well-known MRC methods, AS Reader and AOA Reader, along with four baselines, as a first step towards a BioRead (and BioReadLite) leaderboard. AOA Reader is currently the best method on BioReadLite, with 51.19% test accuracy. Both AOA Reader and AS Reader outperform the baselines by a wide margin on the test subset of BioReadLite. Our re-implementations of the two MRC methods are also publicly available.

MMQA: A Multi-domain Multi-lingual Question-Answering Framework for English and Hindi

Deepak Gupta, Surabhi Kumari, Asif Ekbal and Pushpak Bhattacharyya

In this paper, we assess the challenges for multi-domain, multi-lingual question answering, create necessary resources for benchmarking and develop a baseline model. We curate 500 articles in six different domains from the web. These articles form a comparable corpora of 250 English documents and 250 Hindi documents. From these comparable corpora, we have created 5; 495 question-answer pairs with the questions and answers, both being in English and Hindi. The question can be both factoid or short descriptive types. The answers are categorized in 6 coarse and 63 finer types. To the best of our knowledge, this is the very first attempt towards creating multi-domain, multi-lingual question answering evaluation involving English and Hindi. We develop a deep learning based model for classifying an input question into the coarse and finer categories depending upon the expected answer. Answers are extracted through similarity computation and subsequent ranking. For factoid question, we obtain an MRR value of 49:10% and for short descriptive question, we obtain a BLEU score of 41:37%. Evaluation of question classification model shows the accuracies of 90:12% and 80:30% for coarse and finer classes, respectively.

Session P37 - Social Media Processing (2)

10th May 2018, 14:50

Chair person: **Tetsuro Takahashi** Poster Session

The First 100 Days: A Corpus Of Political Agendas on Twitter

Nathan Green and Septina Larasati

The first 100 days corpus is a curated corpus of the first 100 days of the United States of America's President and the Senate. During the first 100 days, the political parties in the USA try to push their agendas for the upcoming year under the new President. As communication has changed this is primarily being done on Twitter so that the President and Senators can communicate directly with their constituents. We analyzed the current President along with 100 Senators ranging the political spectrum to see the differences in their language usage. The creation of this corpus is intended to help Natural Language Processing (NLP) and Political Science research studying the changing political climate during a shift in power through language. To help accomplish this, the corpus is harvested and normalized in multiple formats. As well, we include gold standard part-of-speech tags for selected individuals including the President. Through analysis of the text, a clear distinction between political parties can be found. This analysis shows the important item of their political agendas during the first 100 days of a new party in power.

Medical Sentiment Analysis using Social Media: Towards building a Patient Assisted System

Shweta Yadav, Asif Ekbal, Sriparna Saha and Pushpak Bhattacharyya

With the enormous growth of Internet, more users have engaged in health communities such as medical forums to gather health-related information, to share experiences about drugs, treatments, diagnosis or to interact with other users with the similar condition in communities. Monitoring social media platforms has recently fascinated medical natural language processing researchers to detect various medical abnormalities such as adverse drug reaction. In this paper, we present a benchmark setup for analyzing the sentiment with respect to users' medical condition considering the information, available in social media in particular. To this end, we have crawled the medical forum

website ‘patient.info’ with opinions about medical condition self-narrated by the users. We constrained ourselves to some of the popular domains such as depression, anxiety, asthma, and allergy. The focus is given on the identification of multiple forms of medical sentiments which can be inferred from users’ medical condition, treatment, and medication. Thereafter, a deep Convolutional Neural Network (CNN) based medical sentiment analysis system is developed for the purpose of evaluation. The resources are made available to the community through LRE map for further research.

An Italian Twitter Corpus of Hate Speech against Immigrants

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti and Marco Stranisci

The paper describes a recently-created Twitter corpus of about 6,000 tweets, annotated for hate speech against immigrants, and developed to be a reference dataset for an automatic system of hate speech monitoring. The annotation scheme was therefore specifically designed to account for the multiplicity of factors that can contribute to the definition of a hate speech notion, and to offer a broader tagset capable of better representing all those factors, which may increase, or rather mitigate, the impact of the message. This resulted in a scheme that includes, besides hate speech, the following categories: aggressiveness, offensiveness, irony, stereotype, and (on an experimental basis) intensity. The paper hereby presented namely focuses on how this annotation scheme was designed and applied to the corpus. In particular, also comparing the annotation produced by CrowdFlower contributors and by expert annotators, we make some remarks about the value of the novel resource as gold standard, which stems from a preliminary qualitative analysis of the annotated data and on future corpus development.

A Large Multilingual and Multi-domain Dataset for Recommender Systems

Giorgia Di Tommaso, Stefano Faralli and Paola Velardi

This paper presents a multi-domain interests dataset to train and test Recommender Systems, and the methodology to create the dataset from Twitter messages in English and Italian. The English dataset includes an average of 90 preferences per user on music, books, movies, celebrities, sport, politics and much more, for about half million users. Preferences are either extracted from messages of users who use Spotify, Goodreads and other similar content sharing platforms, or induced from their ”topical” friends, i.e., followees representing an interest rather than a social relation between peers. In addition, preferred items are matched with Wikipedia articles describing them. This unique feature of our dataset provides a mean to derive a semantic categorization of the preferred items, exploiting available semantic resources linked to Wikipedia such as the Wikipedia Category Graph, DBpedia, BabelNet and others.

RtGender: A Corpus for Studying Differential Responses to Gender

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky and Yulia Tsvetkov

Like many social variables, gender pervasively influences how people communicate with one another. However, prior computational work has largely focused on linguistic gender difference and communications about gender, rather than communications directed to people of that gender, in part due to lack of data. Here, we fill a critical need by introducing a multi-genre corpus of more than 25M comments from five socially and topically diverse sources tagged for the gender of the addressee. Using these data, we describe pilot studies on how differential responses to gender can be measured and analyzed and present 30k annotations for the sentiment and relevance of these responses, showing that across our datasets responses to women are more likely to be emotive and about the speaker as an individual (rather than about the content being responded to). Our dataset enables studying socially important questions like gender bias, and has potential uses for downstream applications such as dialogue systems, gender detection or obfuscation, and debiasing language generation.

A Neural Network Model for Part-Of-Speech Tagging of Social Media Texts

Sara Meftah and Nasredine Semmar

In this paper, we propose a neural network model for Part-Of-Speech (POS) tagging of User-Generated Content (UGC) such as Twitter, Facebook and Web forums. The proposed model is end-to-end and uses both character and word level representations. Character level representations are learned during the training of the model through a Convolutional Neural Network (CNN). For word level representations, we combine several pre-trainned embeddings (Word2Vec, FastText and GloVe). To deal with the issue of the poor availability of annotated social media data, we have implemented a Transfer Learning (TL) approach. We demonstrate the validity and genericity of our model on a POS tagging task by conducting our experiments on five social media languages (English, German, French, Italian and Spanish).

Utilizing Large Twitter Corpora to Create Sentiment Lexica

Valerij Fredriksen, Brage Jahren and Björn Gambäck

The paper describes an automatic Twitter sentiment lexicon creator and a lexicon-based sentiment analysis system. The lexicon creator is based on a Pointwise Mutual Information approach, utilizing 6.25 million automatically labeled tweets and 103 million unlabeled, with the created lexicon consisting of about 3 000 entries. In a comparison experiment, this lexicon beat a manually annotated lexicon. A sentiment analysis system utilizing the created lexicon, and handling both negation and intensification, produces results almost on par with sophisticated machine learning-based systems, while significantly outperforming those in terms of run-time.

Session P38 - Speech Resource/Database (1)

10th May 2018, 14:50

Chair person: **Gilles Adda** Poster Session

The Nautilus Speaker Characterization Corpus: Speech Recordings and Labels of Speaker Characteristics and Voice Descriptions

Laura Fernández Gallardo and Benjamin Weiss

The Nautilus Speaker Characterization corpus is presented. It comprises conversational microphone speech recordings from 300 German speakers (126 males and 174 females) made in 2016/2017 in the acoustically-isolated room Nautilus of the Quality and Usability Lab of the Technische Universität Berlin, Germany. Four scripted and four semi-spontaneous dialogs were elicited from the speakers, simulating telephone call inquiries. Additionally, other spontaneous neutral and emotional speech utterances and questions were produced. Interactions between speakers and their interlocutor (who also conducted the recording session) are provided in separate mono files, accompanied by timestamps and tags that define the speaker's turns. One of the recorded semi-spontanous dialogs has been labeled by external assessors on 34 interpersonal speaker characteristics for each speaker, employing continous sliders. Additionally, 20 selected speakers have been labeled on 34 naive voice descriptions. The corpus labels permit to investigate the speech features that contribute to human perceptions and automatic recognition of speaker social characteristics and interpersonal traits.

Evaluation of Automatic Formant Trackers

Florian Schiel and Thomas Zitzelsberger

Four open source formant trackers, three LPC-based and one based on Deep Learning, were evaluated on the same American English data set VTR-TIMIT. Test data were time-synchronized to avoid differences due to different unvoiced/voiced detection strategies. Default output values of trackers (e.g. producing 500Hz for the first formant, 1500Hz for the second etc.) were filtered from the evaluation data to avoid biased results. Evaluations were performed on the total recording and on three American English vowels [i:], [u] and [] separately. The obtained quality measures showed that all three LPC-based trackers had comparable RSME error results that are about 2 times the inter-labeller error of human labellers. Tracker results were biased considerably (in average too high or low), when the parameter settings of the tracker were not adjusted to the speaker's sex. Deep Learning appeared to outperform LPC-based trackers in general, but not in vowels. Deep Learning has the disadvantage that it requires annotated training material from the same speech domain as the target speech, and a trained Deep Learning tracker is therefore not applicable to other languages.

Design and Development of Speech Corpora for Air Traffic Control Training

Luboš Šmídl, Jan Švec, Daniel Tihelka, Jindrich Matousek, Jan Romportl and Pavel Ircing

The paper describes the process of creation of domain-specific speech corpora containing air traffic control (ATC) communication prompts. Since the ATC domain is highly specific both from the acoustic point-of-view (significant level of noise in the signal, non-native English accents of the speakers, non-standard pronunciation of some frequent words) and the lexical and syntactic perspective (prescribed structure of utterances, rather limited vocabulary), it is useful to collect and annotate data from this specific domain. Actually, the ultimate goal of the research effort of our team was to develop a voice dialogue system simulating the responses of the pilot that could be used for training aspiring air traffic controllers. In order to do so, we needed - among other modules - a domain-specific automatic speech recognition (ASR) and text-to-speech synthesis (TTS) engines. This paper concentrates on the details of the ASR and TTS corpora creation process but also overviews their usage in preparing practical applications and provides links to the distribution channel of the data.

A First South African Corpus of Multilingual Code-switched Soap Opera Speech

Ewald Van der westhuizen and Thomas Niesler

We introduce a speech corpus containing multilingual code-switching compiled from South African soap operas. The corpus contains English, isiZulu, isiXhosa, Setswana and Sesotho speech, paired into four language-balanced subcorpora containing English-isiZulu, English-isiXhosa, English-Setswana and English-Sesotho. In total, the corpus contains 14.3 hours

of annotated and segmented speech. The soap opera speech is typically fast, spontaneous and may express emotion, with a speech rate that is between 1.22 and 1.83 times higher than prompted speech in the same languages. Among the 10343 code-switched utterances in the corpus, 19207 intrasentential language switches are observed. Insertional code-switching with English words is observed to be most frequent. Intraword code-switching, where English words are supplemented with Bantu affixes in an effort to conform to Bantu phonology, is also observed. Most bigrams containing code-switching occur only once, making up between 64% and 92% of such bigrams in each subcorpus.

A Web Service for Pre-segmenting Very Long Transcribed Speech Recordings

Nina Poerner and Florian Schiel

The run time of classical text-to-speech alignment algorithms tends to grow quadratically with the length of the input. This makes it difficult to apply them to very long speech recordings. In this paper, we describe and evaluate two algorithms that pre-segment long recordings into manageable "chunks". The first algorithm is fast but cannot guarantee short chunks on noisy recordings or erroneous transcriptions. The second algorithm reliably delivers short chunks but is less effective in terms of run time and chunk boundary accuracy. We show that both algorithms reduce the run time of the MAUS speech segmentation system to under real-time, even on recordings that could not previously be processed. Evaluation on real-world recordings in three different languages shows that the majority of chunk boundaries obtained with the proposed methods deviate less than 100 ms from a ground truth segmentation. On a separate German studio quality recording, MAUS word segmentation accuracy was slightly improved by both algorithms. The chunking service is freely accessible via a web API in the CLARIN infrastructure, and currently supports 33 languages and dialects.

A Real-life, French-accented Corpus of Air Traffic Control Communications

Estelle Delpech, Marion Laignelet, Christophe Pimm, Céline Raynal, Michal Trzos, Alexandre Arnold and Dominique Pronto

This paper describes the creation of the AIRBUS-ATC corpus, which is a real-life, French-accented speech corpus of Air Traffic Control (ATC) communications (message exchanged between pilots and controllers) intended to build a robust ATC speech recognition engine. The corpus is currently composed of 59 hours of transcribed English audio, along with linguistic and meta-data annotations. It is intended to reach 100 hours by the end of the project. We describe ATC speech specificities, how the audio is collected, transcribed and what techniques were used to ensure transcription quality while limiting transcription costs. A detailed description of the corpus content (speaker gender, accent, role, type of control, speech turn duration) is given. Finally, preliminary results obtained with state-of-the-art speech recognition techniques support the idea that accent-specific corpora will play a pivotal role in building robust ATC speech recognition applications.

Creating Lithuanian and Latvian Speech Corpora from Inaccurately Annotated Web Data

Askars Salimbajevs

This paper describes the method that was used to produce additional acoustic model training data for the less-resourced languages of Lithuanian and Latvian. The method uses existing baseline speech recognition systems for Latvian and Lithuanian to align audio data from the Web with imprecise non-normalised transcripts. From 690 hours of Web data (300h for Latvian, 390h for Lithuanian), we have created additional 378 hours of training data (186h for Latvian and 192 for Lithuanian). Combining this additional data with baseline training data allowed to significantly improve word error rate for Lithuanian from 40% to 23%. Word error rate for the Latvian system was improved from 19% to 17%.

Discovering Canonical Indian English Accents: A Crowdsourcing-based Approach

Sunayana Sitaram, Varun Manjunath, Varun Bharadwaj, Monojit Choudhury, Kalika Bali and Michael Tjalve

Automatic Speech Recognition (ASR) systems typically degrade in performance when recognizing an accent different from the accents in the training data. One way to overcome this problem without training new models for every accent is adaptation. India has over a hundred major languages, which leads to many variants in Indian English accents. Making an ASR system work well for Indian English would involve collecting data for all representative accents in Indian English and then adapting Acoustic Models for each of those accents. However, given the number of languages that exist in India and the lack of a prior work in literature about how many Indian English accents exist, it is difcult to come up with a set of canonical accents that could sufciently capture the variations observed in Indian English. In addition, there is a lack of labeled corpora of accents in Indian English. We approach the problem of determining a set of canonical Indian English accents by taking a crowdsourcing based approach. We conduct a mobile app based user study in which we play audio samples collected from all over India and ask users to identify the geographical origin of the speaker. We measure the consensus among users to come up with a set of candidate accents in Indian English and identify which accents are best recognized and which ones are confusable. We extend our preliminary user study to a web

app-based study that can potentially generate more labeled data for Indian English accents. We describe results and challenges encountered in a pilot study conducted using the web-app and future work to scale up the study.

Extending Search System based on Interactive Visualization for Speech Corpora

Tomoko Ohsuga, Yuichi Ishimoto, Tomoko Kajiyama, Shunsuke Kozawa, Kiyotaka Uchimoto and Shuichi Itahashi

This paper describes a search system that we have developed specifically for speech corpus retrieval. It is difficult for speech corpus users to compare and select suitable corpora from the large number of various language resources in the world. It would be more convenient for users if each data center used a common specification system for describing its corpora. With the "Concentric Ring View (CRV) System" we proposed, users can search for speech corpora interactively and visually by utilizing the attributes peculiar to speech corpora. We have already proposed a set of specification attributes and items as the first step towards standardization, and we have added these attributes and items to the large-scale metadata database "SHACHI", then we connected SHACHI to the CRV system and implemented it as a combined speech corpus search system.

German Radio Interviews: The GRAIN Release of the SFB732 Silver Standard Collection

Katrin Schweitzer, Kerstin Eckart, Markus Gärtner, Agnieszka Falenska, Arndt Riester, Ina Roesiger, Antje Schweitzer, Sabrina Stehwien and Jonas Kuhn

We present GRAIN (German RAdio INterviews) as part of the SFB732 Silver Standard Collection. GRAIN contains German radio interviews and is annotated on multiple linguistic layers. The data has been processed with state-of-the-art tools for text and speech and therefore represents a resource for text-based linguistic research as well as speech science. While there is a gold standard part with manual annotations, the (much larger) silver standard part (which is growing as the radio station releases more interviews) relies completely on automatic annotations. We explicitly release different versions of annotations for the same layers (e.g. morpho-syntax) with the aim to combine and compare multiple layers in order to derive confidence estimations for the annotations. Therefore, parts of the data where the output of several tools match can be considered clear-cut cases, while mismatches hint at areas of interest which are potentially challenging or where rare phenomena can be found.

Preparing Data from Psychotherapy for Natural Language Processing

Margot Mieskes and Andreas Stiegelmayr

Mental health and well-being are growing issues in western civilizations. But at the same time, psychotherapy and further education in psychotherapy is a highly demanding occupation, resulting in a severe gap in patient-centered care. The question which arises from recent developments in natural language processing (NLP) and speech recognition is, how these technologies could be employed to support the therapists in their work and allow for a better treatment of patients. Most research in NLP focuses on analysing the language of patients with various psychological conditions, but only few examples exist that analyse the therapists behavior and the interaction between therapist and patient. We present ongoing work in collecting, preparing and analysing data from psychotherapy sessions together with expert annotations on various qualitative dimensions of these sessions, such as feedback and cooperation. Our aim is to use this data in a classification task, which gives insight into what qualifies for good feedback or cooperation in therapy sessions and employ this information to support psychotherapists in improving the quality of the care they offer.

MirasVoice: A bilingual (English-Persian) speech corpus

Amir Vaheb, Ali Janalizadeh Choobbasti, Mahdi Mortazavi, Saeid Safavi and Behnam Sabeti

Speech and speaker recognition is one of the most important research and development areas and has received quite a lot of attention in recent years. The desire to produce a natural form of communication between humans and machines can be considered the motivating factor behind such developments. Speech has the potential to influence numerous fields of research and development. In this paper, MirasVoice which is a bilingual (English-Farsi) speech corpus is presented. Over 50 native Iranian speakers who were able to speak in both the Farsi and English languages have volunteered to help create this bilingual corpus. The volunteers read text documents and then had to answer questions spontaneously in both English and Farsi. The text-independent GMM-UBM speaker verification engine was designed in this study for validating and exploring the performance of this corpus. This multilingual speech corpus could be used in a variety of language dependent and independent applications. For example, it can be used to investigate the effects of different languages (Farsi and English) on the performance of speaker verification systems. The authors of this paper have

also investigated speaker verification systems performances when using different train/test architectures.

Session O25 - Social Media & Evaluation

10th May 2018, 16:50

Chair person: **Nasredine Semmar** Oral Session

Building an Ellipsis-aware Chinese Dependency Treebank for Web Text

Xuancheng Ren, Xu SUN, Ji Wen, Bingzhen Wei, Weidong Zhan and Zhiyuan Zhang

Web 2.0 has brought with it numerous user-produced data revealing one's thoughts, experiences, and knowledge, which are a great source for many tasks, such as information extraction, and knowledge base construction. However, the colloquial nature of the texts poses new challenges for current natural language processing techniques, which are more adapt to the formal form of the language. Ellipsis is a common linguistic phenomenon that some words are left out as they are understood from the context, especially in oral utterance, hindering the improvement of dependency parsing, which is of great importance for tasks relied on the meaning of the sentence. In order to promote research in this area, we are releasing a Chinese dependency treebank of 319 weibos, containing 572 sentences with omissions restored and contexts reserved.

EuroGames16: Evaluating Change Detection in Online Conversation

Cyril Goutte, Yunli Wang, FangMing Liao, Zachary Zanussi, Samuel Larkin and Yuri Grinberg

We introduce the challenging task of detecting changes from an online conversation. Our goal is to detect significant changes in, for example, sentiment or topic in a stream of messages that are part of an ongoing conversation. Our approach relies on first applying linguistic preprocessing or collecting simple statistics on the messages in the conversation in order to build a time series. Change point detection algorithms are then applied to identify the location of significant changes in the distribution of the underlying time series. We present a collection of sports events on which we can evaluate the performance of our change detection method. Our experiments, using several change point detection algorithms and several types of time series, show that it is possible to detect salient changes in an on-line conversation with relatively high accuracy.

A Deep Neural Network based Approach for Entity Extraction in Code-Mixed Indian Social Media Text

Deepak Gupta, Asif Ekbal and Pushpak Bhattacharyya

The rise in accessibility of web to the masses has led to a spurt in the use of social media making it convenient and powerful way to express and exchange information in their own language(s). India, being enormously diversified country have more than 168 millions users on social media. This diversity is also reflected in their scripts where a majority of users often switch between their native language to be more expressive. These linguistic variations make automatic entity extraction both a necessary and a challenging problem. In this paper, we report our work for entity extraction in a code-mixed environment. Entity extraction is a fundamental component in many natural language processing (NLP) applications. The task of entity extraction faces more challenges while dealing with unstructured and informal texts, and mixing of scripts (i.e., code-mixing) further adds complexities to the process. Our proposed approach is based on the popular deep neural network based Gated Recurrent Unit (GRU) units that discover the higher level features from the text automatically. It does not require handcrafted features or rules, unlike the existing systems. To the best of our knowledge, it is the first attempt for entity extraction from code mixed data using the deep neural network. The proposed system achieves the F-scores of 66.04% and 53.85% for English-Hindi and English-Tamil language pairs, respectively.

PoSTWITA-UD: an Italian Twitter Treebank in Universal Dependencies

Manuela Sanguinetti, Cristina Bosco, Alberto Lavelli, Alessandro Mazzei, Oronzo Antonelli and Fabio Tamburini

Due to the spread of social media-based applications and the challenges posed by the treatment of social media texts in NLP tools, tailored approaches and ad hoc resources are required to provide the proper coverage of specific linguistic phenomena. Various attempts to produce this kind of specialized resources and tools are described in literature. However, most of these attempts mainly focus on PoS-tagged corpora and only a few of them deal with syntactic annotation. This is particularly true for the Italian language, for which such a resource is currently missing. We thus propose the development of PoSTWITA-UD, a collection of tweets annotated according to a well-known dependency-based annotation format: the Universal Dependencies. The goal of this work is manifold, and it mainly consists in creating a resource that, especially for Italian, can be exploited for the training of NLP systems so as to enhance their performance on social media texts. In this paper we focus on the current state of the resource.

Annotating If the Authors of a Tweet are Located at the Locations They Tweet About

Vivek Reddy Doudagiri, Alakananda Vempala and Eduardo Blanco

The locations in a tweet do not always indicate spatial information involving the author of the tweet. In this paper, we investigate whether authors are located or not located in the locations they tweet about, and temporally anchor this spatial information in the tweet timestamp. Specifically, we work with temporal tags centred around the tweet timestamp: longer than 24 hours before or after tweeting, within 24 hours before or after tweeting, and at the time of tweeting. We introduce a corpus of 1,200 location mentions from 1,062 tweets, discuss several annotation samples, and analyze annotator disagreements.

Session O26 - Standards, Validation, Workflows

10th May 2018, 16:50

Chair person: **Yohei Murakami** Oral Session

MOCCA: Measure of Confidence for Corpus Analysis - Automatic Reliability Check of Transcript and Automatic Segmentation

Thomas Kisler and Florian Schiel

The production of speech corpora typically involves manual labor to verify and correct the output of automatic transcription/segmentation processes. This study investigates the possibility of speeding up this correction process using techniques borrowed from automatic speech recognition to predict the location of transcription or segmentation errors in the signal. This was achieved with functionals of features derived from a typical Hidden Markov Model (HMM)-based speech segmentation system and a classification/regression approach based on Support Vector Machine (SVM)/Support Vector Regression (SVR) and Random Forest (RF). Classifiers were tuned in a 10-fold cross validation on an annotated corpus of spontaneous speech. Tests on an independent speech corpus from a different domain showed that transcription errors were predicted with an accuracy of 78% using an SVM, while segmentation errors were predicted in the form of an overlap-measure which showed a Pearson correlation of 0.64 to a ground truth using Support Vector Regression (SVR). The methods described here will be implemented as free-to-use Common Language and Resources and Technology Infrastucture (CLARIN) web services.

Towards an ISO Standard for the Annotation of Quantification

Harry Bunt, James Pustejovsky and Kiyong Lee

This paper presents an approach to the annotation of quantification that is being developed in preparation of the specification of a quantification annotation scheme, as part of an effort of the International Organisation for Standardisation ISO to define interoperable semantic annotation schemes. The paper focuses on the theoretical basis for an ISO standard annotation scheme for quantification phenomena. It is argued that the combination of Generalized Quantifier Theory, neo-Davidsonian event-bases semantics, Discourse Representation Theory, and the ISO Principles of semantic annotation forms a powerful and solid foundation for defining annotations of quantification phenomena with an abstract and a concrete syntax and a compositional semantics. The coverage of the proposed annotation scheme includes both count and mass NP quantifiers, as well as quantification by NPs with syntactically and semantically complex heads with internal quantification and scoping structures, such as inverse linking by prepositional phrases and relative clauses.

Lightweight Grammatical Annotation in the TEI: New Perspectives

Piotr Banski, Susanne Haaf and Martin Mueller

In mid-2017, as part of our activities within the TEI Special Interest Group for Linguists (LingSIG), we submitted to the TEI Technical Council a proposal for a new attribute class that would gather attributes facilitating simple token-level linguistic annotation. With this proposal, we addressed community feedback complaining about the lack of a specific tagset for lightweight linguistic annotation within the TEI. Apart from @lemma and @lemmaRef, up till now TEI encoders could only resort to using the generic attribute @ana for inline linguistic annotation, or to the quite complex system of feature structures for robust linguistic annotation, the latter requiring relatively complex processing even for the most basic types of linguistic features. As a result, there exists now a small set of basic descriptive devices which have been made available at the cost of only very small changes to the TEI tagset. The merit of a predefined TEI tagset for lightweight linguistic annotation is the homogeneity of tagging and thus better interoperability of simple linguistic resources encoded in the TEI. The present paper introduces the new attributes, makes a case for one more addition, and presents the advantages of the new system over the legacy TEI solutions.

A Gold Standard for Multilingual Automatic Term Extraction from Comparable Corpora: Term Structure and Translation Equivalents

Ayla Rigouts Terryn, Veronique Hoste and Els Lefever

Terms are notoriously difficult to identify, both automatically and manually. This complicates the evaluation of the already challenging task of automatic term extraction. With the advent of multilingual automatic term extraction from comparable corpora, accurate evaluation becomes increasingly difficult, since term linking must be evaluated as well as term extraction. A gold standard with manual annotations for a complete comparable corpus has been developed, based on a novel methodology created to accommodate for the intrinsic difficulties of this task. In this contribution, we show how the effort involved in the development of this gold standard resulted, not only in a tool for evaluation, but also in a rich source of information about terms. A detailed analysis of term characteristics illustrates how such knowledge about terms may inspire improvements for automatic term extraction.

Handling Big Data and Sensitive Data Using EUDAT's Generic Execution Framework and the WebLicht Workflow Engine.

Claus Zinn, Wei Qui, Marie Hinrichs, Emanuel Dima and Alexandr Chernov

Web-based tools and workflow engines can often not be applied to data with restrictive property rights and to big data. In both cases, it is better to move the tools to the data rather than having the data travel to the tools. In this paper, we report on the progress to bring together the CLARIN-based WebLicht workflow engine with the EUDAT-based Generic Execution Framework to address this issue.

Session O27 - Treebanks & Parsing

10th May 2018, 16:50

Chair person: **Wenliang Chen** Oral Session

Building a Web-Scale Dependency-Parsed Corpus from CommonCrawl

Alexander Panchenko, Eugen Ruppert, Stefano Faralli, Simone Paolo Ponzetto and Chris Biemann

We present DepCC, the largest-to-date linguistically analyzed corpus in English including 365 million documents, composed of 252 billion tokens and 7.5 billion of named entity occurrences in 14.3 billion sentences from a web-scale crawl of the Common Crawl project. The sentences are processed with a dependency parser and with a named entity tagger and contain provenance information, enabling various applications ranging from training syntax-based word embeddings to open information extraction and question answering. We built an index of all sentences and their linguistic meta-data enabling quick search across the corpus. We demonstrate the utility of this corpus on the verb similarity task by showing that a distributional model trained on our corpus yields better results than models trained on smaller corpora, like Wikipedia. This distributional model outperforms the state of art models of verb similarity trained on smaller corpora on the SimVerb3500 dataset.

Universal Dependencies Version 2 for Japanese

Masayuki Asahara, Hiroshi Kanayama, Takaaki Tanaka, Yusuke Miyao, Sumire Uematsu, Shinsuke Mori, Yuji Matsumoto, Mai Omura and Yugo Murawaki

The Universal Dependencies (UD) project (McDonald et al., 2013) has defined a consistent, crosslinguistic target and syntactic structure representation format. In this presentation, we will show the work of the UD Japanese team. The UD Japanese team was organised by interested people who are developing their own treebanks or parsers. We developed and maintained several UD guidelines (version 2.0) compatible data for Japanese. Most of the data are made through automatic conversion from the existing treebank. The UD annotation guideline was updated from version 1 to version 2 in early 2017. The automatic conversion enabled us to adapt the existing annotation based on traditional Japanese grammar conventions for the UD annotation guideline changes. In this paper, we discuss the current issues of UD Japanese resources until today. These issues come from the difficulty to perform cross-linguistically consistent annotation for the different grammatical system from western European languages. The points at the issues related to the conversions are split into the delimitation (word, phrase and clause), undefined policies of UD guideline, typological systems for UD, and copyright of Japanese language resources.

Developing the Bangla RST Discourse Treebank

Debopam Das and Manfred Stede

We present a corpus development project which builds a corpus in Bangla called the Bangla RST Discourse Treebank. The corpus contains a collection of 266 Bangla text, which are annotated for coherence relations (relations between propositions, such as Cause or Evidence). The texts represent the newspaper genre, which is further divided into eight sub-genres, such as business-related news, editorial columns and sport reports. We use Rhetorical Structure Theory (Mann and Thompson, 1988) as the theoretical framework of the corpus. In particular, we develop our annotation guidelines based on the guidelines used

in the Potsdam Commentary Corpus (Stede, 2016). In the initial phase of the corpus development process, we have annotated 16 texts, and also conducted an inter-annotator agreement study, evaluating the reliability of our guidelines and the reproducibility of our annotation. The corpus upon its completion could be used as a valuable resource for conducting (cross-linguistic) discourse studies for Bangla, and also for developing various NLP applications, such as text summarization, machine translation or sentiment analysis.

A New Version of the Składnica Treebank of Polish Harmonised with the Walenty Valency Dictionary

Marcin Woliński, Elżbieta Hajnicz and Tomasz Bartosiak

This paper reports on developments in the Składnica treebank of Polish which were possible due to the switch to the Walenty valency dictionary. The change required several modifications in the Świgra parser, such as implementing unlike coordination, semantically motivated phrases, and non-standard case values. A procedure to upgrade manually disambiguated trees of Składnica was required as well. Modifications introduced in the treebank included systematic changes of notation and resolving ambiguity between semantically motivated phrases. The procedure of confronting Składnica treebank with the trees generated with the new version of the Świgra parser using Walenty dictionary allowed us to check the consistency of all the resources. This resulted in several corrections introduced in both the treebank and the valence dictionary.

Parse Me if You Can: Artificial Treebanks for Parsing Experiments on Elliptical Constructions

Kira Droganova, Daniel Zeman, Jenna Kanerva and Filip Ginter

In this work we focus on a particular linguistic phenomenon, ellipsis, and explore the latest parsers in order to learn about parsing accuracy and typical errors from the perspective of elliptical constructions. For this purpose we collected and processed outputs of several state-of-the art parsers that took part in the CoNLL 2017 Shared Task. We extended the official shared task evaluation software to obtain focused evaluation of elliptical constructions. Since the studied structures are comparatively rare, and consequently there is not enough data for experimentation, we further describe the creation of a new resource, a semi-artificially constructed treebank of ellipsis.

Session O28 - Morphology & Lexicons

10th May 2018, 16:50

Chair person: **Tamás Varádi** Oral Session

Semi-Automatic Construction of Word-Formation Networks (for Polish and Spanish)

Mateusz Lango, Magda Sevcikova and Zdeněk Žabokrtský

The paper presents a semi-automatic method for the construction of derivational networks. The proposed approach applies a sequential pattern mining technique in order to construct useful morphological features in an unsupervised manner. The features take the form of regular expressions and later are used to feed a machine-learned ranking model. The network is constructed by applying resulting model to sort the lists of possible base words and selecting the most probable ones. This approach, besides relatively small training set and a lexicon, does not require any additional language resources such as a list of alternations groups, POS tags etc. The proposed approach is applied to the lexeme sets of two languages, namely Polish and Spanish, which results in the establishment of two novel word-formation networks. Finally, the network constructed for Polish is merged with the derivational connections extracted from the Polish WordNet and those resulting from the derivational rules developed by a linguist, resulting in the biggest word-formation network for that language. The presented approach is general enough to be adopted for other languages.

A multilingual collection of CoNLL-U-compatible morphological lexicons

Benoît Sagot

We introduce UDLexicons, a multilingual collection of morphological lexicons that follow the guidelines and format of the Universal Dependencies initiative. We describe the three approaches we use to create 53 morphological lexicons covering 38 languages, based on existing resources. These lexicons, which are freely available, have already proven useful for improving part-of-speech tagging accuracy in state-of-the-art architectures.

UniMorph 2.0: Universal Morphology

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner and Mans Hulden

The Universal Morphology (UniMorph) project is a collaborative effort to improve how NLP handles complex morphology across the world's languages. The project releases annotated

morphological data using a universal tagset, the UniMorph schema. Each inflected form is associated with a lemma, which typically carries its underlying lexical meaning, and a bundle of morphological features from our schema. Additional supporting data and tools are also released on a per-language basis when available. UniMorph is based at the Center for Language and Speech Processing (CLSP) at Johns Hopkins University in Baltimore, Maryland. This paper details advances made to the collection, annotation, and dissemination of project resources since the initial UniMorph release described at LREC 2016.

A Computational Architecture for the Morphology of Upper Tanana

Olga Lovick, Christopher Cox, Miikka Silfverberg, Antti Arppe and Mans Hulden

In this paper, we describe a computational model of Upper Tanana, a highly endangered Dene (Athabaskan) language spoken in eastern interior Alaska (USA) and in the Yukon Territory (Canada). This model not only parses and generates Upper Tanana verb forms, but uses the language's verb theme category system, a system of lexical-inflectional verb classes, to additionally predict possible derivations and their morphological behavior. This allows us to model a large portion of the Upper Tanana verb lexicon, making it more accessible to learners and scholars alike. Generated derivations will be compared against the narrative corpus of the language as well to the (much more comprehensive) lexical documentation of closely related languages.

Expanding Abbreviations in a Strongly Inflected Language: Are Morphosyntactic Tags Sufficient?

Piotr Żelasko

In this paper, the problem of recovery of morphological information lost in abbreviated forms is addressed with a focus on highly inflected languages. Evidence is presented that the correct inflected form of an expanded abbreviation can in many cases be deduced solely from the morphosyntactic tags of the context. The prediction model is a deep bidirectional LSTM network with tag embedding. The training and evaluation data are gathered by finding the words which could have been abbreviated and using their corresponding morphosyntactic tags as the labels, while the tags of the context words are used as the input features for classification. The network is trained on over 10 million words from the Polish Sejm Corpus and achieves 74.2% prediction accuracy on a smaller, but more general National Corpus of Polish. The analysis of errors suggests that performance in this task may improve if some prior knowledge about the abbreviated word is incorporated into the model.

Session P39 - Conversational Systems/Dialogue/Chatbots/Human-Robot Interaction (2)

10th May 2018, 16:50

Chair person: **Johannes Kraus** Poster Session

Dialog Intent Structure: A Hierarchical Schema of Linked Dialog Acts

Silvia Pareti and Tatiana Lando

In this paper, we present a new hierarchical and extensible schema for dialog representation. The schema captures the pragmatic intents of the conversation independently from any semantic representation. This schema was developed to support computational applications, be applicable to different types of dialogs and domains and enable large-scale non-expert annotation. The schema models dialog as a structure of linked units of intent, dialog acts, that are annotated on minimal spans of text, functional segments. Furthermore, we categorise dialog acts based on whether they express a primary or secondary intent and whether the intent is explicit or implicit. We successfully tested the schema on an heterogeneous corpus of human-human dialogs comprising both spoken and chat interactions.

JDCFC: A Japanese Dialogue Corpus with Feature Changes

Tetsuaki Nakamura and Daisuke Kawahara

In recent years, the importance of dialogue understanding systems has been increasing. However, it is difficult for computers to deeply understand our daily conversations because we frequently use emotional expressions in conversations. This is partially because there are no large-scale corpora focusing on the detailed relationships between emotions and utterances. In this paper, we propose a dialogue corpus constructed based on our knowledge base, called the Japanese Feature Change Knowledge Base (JFCKB). In JFCKB and the proposed corpus, the feature changes (mainly emotions) of arguments in event sentences (or utterances) and those of the event sentence recognizers (or utterance recognizers) are associated with the event sentences (or utterances). The feature change information of arguments in utterances and those of the utterance recognizers, replies to the utterances, and the reasonableness of the replies were gathered through crowdsourcing tasks. We conducted an experiment to investigate whether a machine learning method can recognize the reasonableness of a given conversation. Experimental result suggested the usefulness of our proposed corpus.

Japanese Dialogue Corpus of Information Navigation and Attentive Listening Annotated with Extended ISO-24617-2 Dialogue Act Tags

Koichiro Yoshino, Hiroki Tanaka, Kyoshiro Sugiyama, Makoto Kondo and Satoshi Nakamura

Large-scale dialogue data annotated with dialogue states is necessary to model a natural conversation with machines. However, large-scale conventional dialogue corpora are mainly built for specified tasks (e.g., task-oriented systems for restaurant or bus information navigation) with specially designed dialogue states. Text-chat based dialogue corpora have also been built due to the growth of social communication through the internet; however, most of them do not reflect dialogue behaviors in face-to-face conversation, including backchannelings or interruptions. In this paper, we try to build a corpus that covers a wider range of dialogue tasks than existing task-oriented systems or text-chat systems, by transcribing face-to-face dialogues held in natural conversational situations in tasks of information navigation and attentive listening. The corpus is recorded in Japanese and annotated with an extended ISO-24617-2 dialogue act tag-set, which is defined to see behaviors in natural conversation. The developed data can be used to build a dialogue model based on the ISO-24617-2 dialogue act tags.\\

The Niki and Julie Corpus: Collaborative Multimodal Dialogues between Humans, Robots, and Virtual Agents

Ron Artstein, Jill Boberg, Alesia Gainer, Jonathan Gratch, Emmanuel Johnson, Anton Leuski, Gale Lucas and David Traum

The Niki and Julie corpus contains more than 600 dialogues between human participants and a human-controlled robot or virtual agent, engaged in a series of collaborative item-ranking tasks designed to measure influence. Some of the dialogues contain deliberate conversational errors by the robot, designed to simulate the kinds of conversational breakdown that are typical of present-day automated agents. Data collected include audio and video recordings, the results of the ranking tasks, and questionnaire responses; some of the recordings have been transcribed and annotated for verbal and nonverbal feedback. The corpus has been used to study influence and grounding in dialogue. All the dialogues are in American English.

Constructing a Chinese Medical Conversation Corpus Annotated with Conversational Structures and Actions

Nan Wang, Yan Song and Fei Xia

Overuse of antibiotics and the attributed bacterial resistance is one of the most serious global public health crises today. Previous research reported that patients' advocacy for antibiotic treatment was consequential on antibiotic over-prescribing. To investigate how the advocacy and other factors contribute to antibiotic over-prescribing, qualitative and quantitative analysis of doctor-patient conversation can yield valuable findings. In this paper, we introduce AMed (Annotated Corpus of Medical Conversations), a manually transcribed corpus of medical dialogue in Chinese pediatric consultations, with annotation of conversational structures and actions. Based on the annotation, a significant association between patient request for antibiotic and antibiotic over-prescribing is discovered. As this corpus is the first with annotation of conversational structures and actions on medical consultation conversations in Chinese, it can be a valuable resource for discourse and dialogue research in general, and for the understanding of human collaboration and negotiation behavior in clinical consultations in particular. Furthermore, findings from analyses of the corpus can shed light on ways to improve physician-patient communication in order to reduce antibiotic over-prescribing.

Predicting Nods by using Dialogue Acts in Dialogue

Ryo Ishii, Ryuichiro Higashinaka and Junji Tomita

In addition to verbal behavior, nonverbal behavior is an important aspect for an embodied dialogue system to be able to conduct a smooth conversation with the user. Researchers have focused on automatically generating nonverbal behavior from speech and language information of dialogue systems. We propose a model to generate head nods accompanying utterance from natural language. To the best of our knowledge, previous studies generated nods from the final morphemes at the end of an utterance. In this study, we focused on dialog act information indicating the intention of an utterance and determined whether this information is effective for generating nods. First, we compiled a Japanese corpus of 24 dialogues including utterance and nod information. Next, using the corpus, we created a model that estimates whether a nod occurs during an utterance by using a morpheme at the end of a speech and dialog act. The results show that our estimation model incorporating dialog acts outperformed a model using morpheme information. The results suggest that dialog acts have the potential to be a strong predictor with which to generate nods automatically.

Modeling Collaborative Multimodal Behavior in Group Dialogues: The MULTISIMO Corpus

Maria Koutsombogera and Carl Vogel

We present a multimodal corpus that has been recently developed within the MULTISIMO project and targets the investigation and modeling of collaborative aspects of multimodal behavior in

groups that perform simple tasks. The corpus consists of a set of human-human interactions recorded in multiple modalities. In each interactive session two participants collaborate with each other to solve a quiz while assisted by a facilitator. The corpus has been transcribed and annotated with information related to verbal and non-verbal signals. A set of additional annotation and processing tasks are currently in progress. The corpus includes survey materials, i.e. personality tests and experience assessment questionnaires filled in by all participants. This dataset addresses multiparty collaborative interactions and aims at providing tools for measuring collaboration and task success based on the integration of the related multimodal information and the personality traits of the participants, but also at modeling the multimodal strategies that members of a group employ to discuss and collaborate with each other. The corpus is designed for public release.

A Semi-autonomous System for Creating a Human-Machine Interaction Corpus in Virtual Reality: Application to the ACORFORMed System for Training Doctors to Break Bad News

Magalie Ochs, Philippe Blache, Grégoire De Montcheuil, Jean-Marie Pergandi, Jorane Saubesty, Daniel Francon and Daniel Mestre

In this paper, we introduce a two-step corpora-based methodology, starting from a corpus of human-human interactions to construct a semi-autonomous system in order to collect a new corpus of human-machine interaction, a step before the development of a fully autonomous system constructed based on the analysis of the collected corpora. The presented methodology is illustrated in the context of a virtual reality training platform for doctors breaking bad news.

Construction of English-French Multimodal Affective Conversational Corpus from TV Dramas

Sashi Novitasari, Quoc Truong Do, Sakriani Sakti, Dessi Lestari and Satoshi Nakamura

Recently, there has been an increase of interest in constructing corpora containing social-affective interactions. But the availability of multimodal, multilingual, and emotionally rich corpora remains limited. The tasks of recording and transcribing actual human-to-human affective conversations are also tedious and time-consuming. This paper describes construction of a multimodal affective conversational corpus based on TV dramas. The data contain parallel English-French languages in lexical, acoustic, and facial features. In addition, we annotated the part of the English data with speaker and emotion information. Our corpus can be utilized to develop and assess such tasks as speaker and emotion recognition, affective speech recognition and synthesis, linguistic, and paralinguistic speech-to-speech translation as well as a multimodal dialog system.

QUEST: A Natural Language Interface to Relational Databases

Vadim Sheinin, Elahe Khorasani, Hangu Yeo, Kun Xu, Ngoc Phuoc An Vo and Octavian Popescu

Natural language interfaces to databases systems allow the user to use natural language to interrogate a database. Current systems mainly focus on simple queries but neglect nested queries, which are predominant in real cases. We present a NLIDB system, QUEST, which is able to cope with nested logic queries, without imposing any restriction on the input query. QUEST outperforms a strong baseline system by 11% accuracy.

Session P40 - Language Modelling

10th May 2018, 16:50

Chair person: **Bolette Pedersen** Poster Session

TF-LM: TensorFlow-based Language Modeling Toolkit

Lyan Verwimp, Hugo Van hamme and Patrick Wambacq

Recently, an abundance of deep learning toolkits has been made freely available. These toolkits typically offer the building blocks and sometimes simple example scripts, but designing and training a model still takes a considerable amount of time and knowledge. We present language modeling scripts based on TensorFlow that allow one to train and test competitive models directly, by using a pre-defined configuration or changing it to their needs. There are several options for input features (words, characters, words combined with characters, character n-grams) and for batching (sentence- or discourse-level). The models can be used to test the perplexity, predict the next word(s), re-score hypotheses or generate debugging files for interpolation with n-gram models. Additionally, we make available LSTM language models trained on a variety of Dutch texts and English benchmarks, that can be used immediately, thereby avoiding the time and computationally expensive training process. The toolkit is open source and can be found at https://github.com/lverwimp/tf-lm.

Grapheme-level Awareness in Word Embeddings for Morphologically Rich Languages

Suzi Park and Hyopil Shin

Learning word vectors from character level is an effective method to improve word embeddings for morphologically rich languages. However, most of these techniques have been applied to languages that are inflectional and written in Roman alphabets. In this paper, we investigate languages that are agglutinative and represented by non-alphabetic scripts, choosing Korean as a

case study. We present a grapheme-level coding procedure for neural word embedding that utilizes word-internal features that are composed of syllable characters (Character CNN). Observing that our grapheme-level model is more capable of representing functional and semantic similarities, grouping allomorphs, and disambiguating homographs than syllable-level and word-level models, we recognize the importance of knowledge on the morphological typology and diversity of writing systems.

Building a Constraint Grammar Parser for Plains Cree Verbs and Arguments

Katherine Schmirler, Antti Arppe, Trond Trosterud and Lene Antonsen

This paper discusses the development and application of a Constraint Grammar parser for the Plains Cree language. The focus of this parser is the identification of relationships between verbs and arguments. The rich morphology and non-configurational syntax of Plains Cree make it an excellent candidate for the application of a Constraint Grammar parser, which is comprised of sets of constraints with two aims: 1) the disambiguation of ambiguous word forms, and 2) the mapping of syntactic relationships between word forms on the basis of morphological features and sentential context. Syntactic modelling of verb and argument relationships in Plains Cree is demonstrated to be a straightforward process, though various semantic and pragmatic features should improve the current parser considerably. When applied to even a relatively small corpus of Plains Cree, the Constraint Grammar parser allows for the identification of common word order patterns and for relationships between word order and information structure to become apparent.

BPEmb: Tokenization-free Pre-trained Subword Embeddings in 275 Languages

Benjamin Heinzerling and Michael Strube

We present BPEmb, a collection of pre-trained subword unit embeddings in 275 languages, based on Byte-Pair Encoding (BPE). In an evaluation using fine-grained entity typing as testbed, BPEmb performs competitively, and for some languages better than alternative subword approaches, while requiring vastly fewer resources and no tokenization. BPEmb is available at https://github.com/bheinzerling/bpemb.

Session P41 - Natural Language Generation

10th May 2018, 16:50

Chair person: **Ineke Schuurman** Poster Session

Reference production in human-computer interaction: Issues for Corpus-based Referring Expression Generation

Danillo Rocha and Ivandré Paraboni

In the Natural Language Generation field, Referring Expression Generation (REG) studies often make use of experiments involving human subjects for the collection of corpora of definite descriptions. Experiments of this kind usually make use of web-based settings in which a single subject acts as a speaker with no particular addressee in mind (as a kind of monologue situation), or in which participant pairs are engaged in an actual dialogue. Both so-called monologue and dialogue settings are of course instances of real language use, but it is not entirely clear whether these situations are truly comparable or, to be more precise, whether REG studies may draw conclusions regarding attribute selection, referential overspecification and others regardless of the mode of communication. To shed light on this issue, in this work we developed a parallel, semantically annotated corpus of monologue and dialogue referring expressions, and carried out an experiment to compare instances produced in both modes of communication. Preliminary results suggest that human reference production may be indeed affected by the presence of a second (specific) human participant as the receiver of the communication in a number of ways, an observation that may be relevant for studies in REG and related fields.

Definite Description Lexical Choice: taking Speaker's Personality into account

Alex Lan and Ivandré Paraboni

In Natural Language Generation (NLG), Referring Expression Generation (REG) lexical choice is the subtask that provides words to express a given input meaning representation. Since lexical choices made in real language use tend to vary greatly across speakers, computational models of lexicalisation have long addressed the issue of human variation in the REG field as well. However, studies of this kind will often rely on large collections of pre-recorded linguistic examples produced by every single speaker of interest, and on every domain under consideration, to obtain meaning-to-text mappings from which the lexicalisation model is built. As a result, speaker-dependent lexicalisation may be impractical when suitable annotated corpora are not available. As an alternative to corpus-based approaches of this kind, this paper argues that differences across human speakers

may be at least partially influenced by personality, and presents a personality-dependent lexical choice model for REG that is, to the best of our knowledge, the first of its kind. Preliminary results show that our personality-dependent approach outperforms a standard lexicalisation model (i.e., based on meaning-to-text mappings alone), and that the use of personality information may be a viable alternative to strategies that rely on corpus knowledge.

Referring Expression Generation in time-constrained communication

André Mariotti and Ivandré Paraboni

In game-like applications and many others, an underlying Natural Language Generation system may have to express urgency or other dynamic aspects of a fast-evolving situation as text, which may be considerably different from text produced under so-called 'normal' circumstances (e.g., without time constrains). As a means to shed light on possible differences of this kind, this paper addresses the computational generation of natural language text in time-constrained communication by presenting two experiments that use the attribute selection task of definite descriptions (or Referring Expression Generation - REG) as a working example. In the first experiment, we describe a psycholinguistic study in which human participants are engaged in a time-constrained reference production task. This results in a corpus of time-constrained descriptions to be compared with 'normal' descriptions available from an existing (i.e., with no time constraint) REG corpus. In the second experiment, we discuss how a REG algorithm may be customised so as to produce time-constrained descriptions that resemble those produced by human speakers in similar situations. The proposed algorithm is then evaluated against the time-constrained descriptions produced by the human subjects in the first experiment, and it is shown to outperform standard approaches to REG in these conditions.

Incorporating Semantic Attention in Video Description Generation

Natsuda Laokulrat, Naoaki Okazaki and Hideki Nakayama

Automatically generating video description is one of the approaches to enable computers to deeply understand videos, which can have a great impact and can be useful to many other applications. However, generated descriptions by computers often fail to correctly mention objects and actions appearing in the videos. This work aims to alleviate this problem by including external fine-grained visual information, which can be detected from all video frames, in the description generation model. In this paper, we propose an LSTM-based sequence-to-sequence model with semantic attention mechanism for video description generation. The model is flexible so that we can change the source of the external information without affecting the encoding and decoding parts of the model. The results show that using semantic attention to selectively focus on external fine-grained visual information can guide the system to correctly mention objects and actions in videos and have a better quality of video descriptions.

GenDR: A Generic Deep Realizer with Complex Lexicalization

François Lareau, Florie Lambrey, Ieva Dubinskaite, Daniel Galarreta-Piquette and Maryam Nejat

We present a generic deep realizer called GenDR, which takes as input an abstract semantic representation of predicate-argument relations, and produces corresponding syntactic dependency structures in English, French, Lithuanian and Persian, with the possibility to fairly easily add more languages. It is generic in that it is designed to operate across a wide range of languages and applications, given the appropriate lexical resources. The focus is on the lexicalization of multiword expressions, with built-in rules to handle thousands of different cross-linguistic patterns of collocations (intensifiers, support verbs, causatives, etc.), and on rich paraphrasing, with the ability to produce many syntactically and lexically varied outputs from the same input. The system runs on a graph transducer, MATE, and its grammar design is directly borrowed from MARQUIS, which we have trimmed down to its core and built upon. The grammar and demo dictionaries are distributed under a CC-BY-SA licence (http://bit.ly/2x8xGVO). This paper explains the design of the grammar, how multiword expressions (especially collocations) are dealt with, and how the syntactic structure is derived from the relative communicative salience of the meanings involved.

A Detailed Evaluation of Neural Sequence-to-Sequence Models for In-domain and Cross-domain Text Simplification

Sanja Štajner and Sergiu Nisioi

We present a detailed evaluation and analysis of neural sequence-to-sequence models for text simplification on two distinct datasets: Simple Wikipedia and Newsela. We employ both human and automatic evaluation to investigate the capacity of neural models to generalize across corpora, and we highlight challenges that these models face when tested on a different genre. Furthermore, we establish a strong baseline on the Newsela dataset and show that a simple neural architecture can be efficiently used for in-domain and cross-domain text simplification.

Don't Annotate, but Validate: a Data-to-Text Method for Capturing Event Data

Piek Vossen, Filip Ilievski, Marten Postma and Roxane Segers

In this paper, we present a new method to obtain large volumes of high-quality text corpora with event data for studying identity and reference relations. We report on the current methods to create event reference data by annotating texts and deriving the event data a posteriori. Our method starts from event registries in which event data is defined a priori. From this data, we extract so-called Microworlds of referential data with the Reference Texts that report on these events. This makes it possible to establish referential relations with high precision easily and at a large scale. In a pilot, we successfully obtained data from these resources with extreme ambiguity and variation, while maintaining the identity and reference relations and without having to annotate large quantities of texts word-by-word. The data from this pilot was annotated using an annotation tool created specifically in order to validate our method and to enrich the reference texts with event coreference annotations. This annotation process resulted in the Gun Violence Corpus, whose development process and outcome are described in this paper.

RDF2PT: Generating Brazilian Portuguese Texts from RDF Data

Diego Moussallem, Thiago Ferreira, Marcos Zampieri, Maria Cláudia Cavalcanti, Geraldo Xexéo, Mariana Neves and Axel-Cyrille Ngonga Ngomo

The generation of natural language from RDF data has recently gained significant attention due to the continuous growth of Linked Data. A number of these approaches generate natural language in languages other than English, however, no work has been proposed to generate Brazilian Portuguese texts out of RDF. We address this research gap by presenting RDF2PT, an approach that verbalizes RDF data to Brazilian Portuguese language. We evaluated RDF2PT in an open questionnaire with 44 native speakers divided into experts and non-experts. Our results suggest that RDF2PT is able to generate text which is similar to that generated by humans and can hence be easily understood.

Towards a music-language mapping

Michele Berlingerio and Francesca Bonin

We explore a novel research idea, that we call Musical Language Processing (MLP), which investigates the possibility of a musical input to speech interaction systems. We present the first attempts at finding a mapping between musical pieces and dialogues, based on the frequency of musical patterns. Our findings on one possible alignment between classical piano compositions and dialogues from popular TV series are encouraging, and open the way to further investigations along this line of research.

Up-cycling Data for Natural Language Generation

Amy Isard, Jon Oberlander and Claire Grover

Museums and other cultural heritage institutions have large databases of information about the objects in their collections, and existing Natural Language Generation (NLG) systems can generate fluent and adaptive texts for visitors, given appropriate input data, but there is typically a large amount of expert human effort required to bridge the gap between the available and the required data. We describe automatic processes which aim to significantly reduce the need for expert input during the conversion and up-cycling process. We detail domain-independent techniques for processing and enhancing data into a format which allows an existing NLG system to create adaptive texts. First we normalize the dates and names which occur in the data, and we link to the Semantic Web to add extra object descriptions. Then we use Semantic Web queries combined with a wide coverage grammar of English to extract relations which can be used to express the content of database fields in language accessible to a general user. As our test domain we use a database from the Edinburgh Musical Instrument Museum.

Session P42 - Semantics (2)

10th May 2018, 16:50

Chair person: **Cécile Fabre** Poster Session

Neural Models of Selectional Preferences for Implicit Semantic Role Labeling

Minh Le and Antske Fokkens

Implicit Semantic Role Labeling is a challenging task: it requires high-level understanding of the text while annotated data is very limited. Due to the lack of training data, most researches either resort to simplistic machine learning methods or focus on automatically acquiring training data. In this paper, we explore the possibilities of using more complex and expressive machine learning models trained on a large amount of explicit roles. In addition, we compare the impact of one-way and multi-way selectional preference with the hypothesis that the added information in multi-way models are beneficial. Although our models surpass a baseline that uses prototypical vectors for SemEval-2010, we otherwise face mostly negative results. Selectional preference models perform lower than the baseline on ON5V, a dataset of five ambiguous and frequent verbs. They are also outperformed by the Na ıve Bayes model of Feizabadi

and Pado (2015) on both datasets. Even though multi-way selectional preference improves results for predicting explicit semantic roles compared to one-way selectional preference, it harms performance for implicit roles. We release our source code, including the reimplementation of two previously unavailable systems to enable further experimentation.

A database of German definitory contexts from selected web sources

Adrien Barbaresi, Lothar Lemnitzer and Alexander Geyken

We introduce work on the detection of definitory contexts designed to speed up two lexicographical tasks: searching for the exact meaning(s) of terms and providing usable input for paraphrasing. Our database is built from a specialized web corpus using a robust pattern-based extraction method. The corresponding interface displays information for a large range of lexical units. The contributions of this article are threefold: we describe both acquisition and extraction, provide a qualitative assessment of the method, and present an interface to access the data.

Annotating Abstract Meaning Representations for Spanish

Noelia Migueles-Abraira, Rodrigo Agerri and Arantza Diaz de Ilarraza

Until recently, semantic annotations for different semantic phenomena were independent and unconnected. The Abstract Meaning Representation (AMR) project arised out of the need to create a broad-coverage semantic bank containing a unified set of semantic information represented in simple, single-rooted, easy-to-read structures. Because the semantic representation language proposed in AMR is biased towards English, annotating AMR structures for other languages, such as Spanish, is not a trivial task. In this paper we propose a linguistic method that we believe would help lay the groundwork for building a large semantic bank for Spanish and would guide those who would like to implement it for other languages. Thus, we analyze a broad spectrum of Spanish linguistic phenomena to come up with suggestions to adapt the current guidelines so that it is possible to annotate AMRs for Spanish. As a result of this work, we make available the first public online repository containing manually annotated Spanish AMRs.

Browsing the Terminological Structure of a Specialized Domain: A Method Based on Lexical Functions and their Classification

Marie-Claude L' Homme, Benoît Robichaud and Nathalie Prévil

This paper describes a method for browsing relations between terms and unveiling the terminological structure of a specialized domain. The method consists in expanding a graph that takes as input the relations encoded in a multilingual terminological resource called the DiCoEnviro that contains terms in the field of the environment. In the DiCoEnviro, terminological relations are encoded using lexical functions (Melčuk et al. 1995) and further classified in families defined on the basis of the properties of relations. We seek to provide users with an explicit and intuitive representation of a wide variety of relations. We also make the most of the richness of the encoding, while implementing some graphical choices to make their interpretation as clear as possible for end users. The method is implemented in a tool called NeoVisual that provides access to more than 11,000 relations in English and 15,000 relations in French. Portuguese is also included and coverage in all languages will increase as new entries are added to the DiCoEnviro.

Rollenwechsel-English: a large-scale semantic role corpus

Asad Sayeed, Pavel Shkadzko and Vera Demberg

We present the Rollenwechsel-English (RW-eng) corpus, a large corpus of automatically-labelled semantic frames extracted from the ukWaC corpus and BNC using Propbank roles. RW-eng contains both full-phrase constituents for labelled roles as well as heads identified by a series of heuristics. This corpus is of a scale and size suitable for new deep learning approaches to language modelling and distributional semantics, particularly as it pertains to generalized event knowledge. We describe the structure of this corpus, tools for its use, and successful use cases.

Towards a Standardized Dataset for Noun Compound Interpretation

Girishkumar Ponkiya, Kevin Patel, Pushpak Bhattacharyya and Girish K. Palshikar

Noun compounds are interesting constructs in Natural Language Processing (NLP). Interpretation of noun compounds is the task of uncovering a relationship between component nouns of a noun compound. There has not been much progress in this field due to lack of a standardized set of relation inventory and associated annotated dataset which can be used to evaluate suggested solutions. Available datasets in the literature suffer from two problems. Firstly, the approaches to creating some of

the relation inventories and datasets are statistically motivated, rather than being linguistically motivated. Secondly, there is little overlap among the semantic relation inventories used by them. We attempt to bridge this gap through our paper. We present a dataset that is (a) linguistically grounded by using Levi (1978)'s theory, and (b) uses frame elements of FrameNet as its semantic relation inventory. The dataset consists of 2,600 examples created by an automated extraction from FrameNet annotated corpus, followed by a manual investigation. These attributes make our dataset useful for noun compound interpretation in a general-purpose setting.

Structured Interpretation of Temporal Relations

Yuchen Zhang and Nianwen Xue

Temporal relations between events and time expressions in a document are often modeled in an unstructured manner where relations between individual pairs of time expressions and events are considered in isolation. This often results in inconsistent and incomplete annotation and computational modeling. We propose a novel annotation approach where events and time expressions in a document form a dependency tree in which each dependency relation corresponds to an instance of temporal anaphora where the antecedent is the parent and the anaphor is the child. We annotate a corpus of 235 documents using this approach in the two genres of news and narratives, with 48 documents doubly annotated. We report a stable and high inter-annotator agreement on the doubly annotated subset, validating our approach, and perform a quantitative comparison between the two genres of the entire corpus. We make this corpus publicly available.

NL2Bash: A Corpus and Semantic Parser for Natural Language Interface to the Linux Operating System

Xi Victoria Lin, Chenglong Wang, Luke Zettlemoyer and Michael D. Ernst

We present new data and semantic parsing methods for the problem of mapping English sentences to Bash commands (NL2Bash). Our long-term goal is to enable any user to perform otherwise repetitive computer operations (such as file manipulation, search, and application-specific scripting) by simply stating their goals in English. We take a first step in this domain, by providing a large new dataset of challenging but commonly used Bash commands and expert-written English descriptions, along with the baseline methods to establish performance levels on this task.

World Knowledge for Abstract Meaning Representation Parsing

Charles Welch, Jonathan K. Kummerfeld, Song Feng and Rada Mihalcea

World Knowledge for Abstract Meaning Representation Parsing Charles Welch, Jonathan K. Kummerfeld, Song Feng, Rada Mihalcea In this paper we explore the role played by world knowledge in semantic parsing. We look at the types of errors that currently exist in a state-of-the-art Abstract Meaning Representation (AMR) parser, and explore the problem of how to integrate world knowledge to reduce these errors. We look at three types of knowledge from (1) WordNet hypernyms and super senses, (2) Wikipedia entity links, and (3) retraining a named entity recognizer to identify concepts in AMR. The retrained entity recognizer is not perfect and cannot recognize all concepts in AMR and we examine the limitations of the named entity features using a set of oracles. The oracles show how performance increases if it can recognize different subsets of AMR concepts. These results show improvement on multiple fine-grained metrics, including a 6% increase in named entity F-score, and provide insight into the potential of world knowledge for future work in Abstract Meaning Representation parsing.

Session P43 - Speech Processing

10th May 2018, 16:50

Chair person: **Sebastian Stüker** Poster Session

Improved Transcription and Indexing of Oral History Interviews for Digital Humanities Research

Michael Gref, Joachim Köhler and Almut Leh

This paper describes different approaches to improve the transcription and indexing quality of the Fraunhofer IAIS Audio Mining system on Oral History interviews for the Digital Humanities Research. As an essential component of the Audio Mining system, automatic speech recognition faces a lot of difficult challenges when processing Oral History interviews. We aim to overcome these challenges using state-of-the-art automatic speech recognition technology. Different acoustic modeling techniques, like multi-condition training and sophisticated neural networks, are applied to train robust acoustic models. To evaluate the performance of these models on Oral History interviews a German Oral History test-set is presented. This test-set represents the large audio-visual archives "Deutsches Gedächtnis" of the Institute for History and Biography. The combination of the different applied techniques results in a word error rate reduced by 28.3% relative on this test-set compared to the current baseline

system while only one eighth of the previous amount of training data is used. In context of these experiments new opportunities are set out for Oral History research offered by Audio Mining. Also the workflow is described used by Audio Mining to process long audio-files to automatically create time-aligned transcriptions.

Sound Signal Processing with Seq2Tree Network

Weicheng Ma, Kai Cao, Zhaoheng Ni, Peter Chin and Xiang Li

Long Short-Term Memory (LSTM) and its variants have been the standard solution to sequential data processing tasks because of their ability to preserve previous information weighted on distance. This feature provides the LSTM family with additional information in predictions, compared to regular Recurrent Neural Networks (RNNs) and Bag-of-Words (BOW) models. In other words, LSTM networks assume the data to be chain-structured. The longer the distance between two data points, the less related the data points are. However, this is usually not the case for real multimedia signals including text, sound and music. In real data, this chain-structured dependency exists only across meaningful groups of data units but not over single units directly. For example, in a prediction task over sound signals, a meaningful word could give a strong hint to its following word as a whole but not the first phoneme of that word. This undermines the ability of LSTM networks in modeling multimedia data, which is pattern-rich. In this paper we take advantage of Seq2Tree network, a dynamically extensible tree-structured neural network architecture which helps solve the problem LSTM networks face in sound signal processing tasks-the unbalanced connections among data units inside and outside semantic groups. Experiments show that Seq2Tree network outperforms the state-of-the-art Bidirectional LSTM (BLSTM) model on a signal and noise separation task (CHiME Speech Separation and Recognition Challenge).

Open ASR for Icelandic: Resources and a Baseline System

Anna Björk Nikulásdóttir, Inga Rún Helgadóttir, Matthías Pétursson and Jón Guðnason

Developing language resources is an important task when creating a speech recognition system for a less-resourced language. In this paper we describe available language resources and their preparation for use in a large vocabulary speech recognition (LVSR) system for Icelandic. The content of a speech corpus is analysed and training and test sets compiled, a pronunciation dictionary is extended, and text normalization for language modeling performed. An ASR system based on neural networks is implemented using these resources and tested using different acoustic training sets. Experimental results show a clear increase in word-error-rate (WER) when using smaller training sets, indicating that extension of the speech corpus for training would improve the system. When testing on data with known vocabulary only, the WER is 7.99%, but on an open vocabulary test set the WER is 15.72%. Furthermore, impact of the content of the acoustic training corpus is examined. The current results indicate that an ASR system could profit from carefully selected phonotactical data, however, further experiments are needed to verify this impression. The language resources are available on http://malfong.is and the source code of the project can be found on https://github.com/cadia-lvl/ice-asr/tree/master/ice-kaldi.

Towards Neural Speaker Modeling in Multi-Party Conversation: The Task, Dataset, and Models

Zhao Meng, Lili Mou and Zhi Jin

Neural network-based dialog systems are attracting increasing attention in both academia and industry. Recently, researchers have begun to realize the importance of speaker modeling in neural dialog systems, but there lacks established tasks and datasets. In this paper, we propose speaker classification}as a surrogate task for general speaker modeling, and collect massive data to facilitate research in this direction. We further investigate temporal-based and content-based models of speakers, and propose several hybrids of them. Experiments show that speaker classification is feasible, and that hybrid models outperform each single component.

Discriminating between Similar Languages on Imbalanced Conversational Texts

Junqing He, Xian Huang, Xuemin Zhao, Yan Zhang and Yonghong Yan

Discriminating between similar languages (DSL) on conversational texts is a challenging task. This paper aims at discriminating between limited-resource languages on short conversational texts, like Uyghur and Kazakh. Considering that Uyghur and Kazakh data are severely imbalanced, we leverage an effective compensation strategy to build a balanced Uyghur and Kazakh corpus. Then we construct a maximum entropy classifier based on morphological features to discriminate between the two languages and investigate the contribution of each feature. Empirical results suggest that our system achieves an accuracy of 95.7\% on our Uyghur and Kazakh dataset, which is higher than that of the CNN classifier. We also apply our system to the out-of-domain subtask of VarDial' 2016 DSL shared tasks to test the system's performance on short conversational texts of other similar languages. Though with much less preprocessing, our system outperforms the champions on both test sets B1 and B2.

Data-Driven Pronunciation Modeling of Swiss German Dialectal Speech for Automatic Speech Recognition

Michael Stadtschnitzer and Christoph Schmidt

Automatic speech recognition is a requested technique in many fields like automatic subtitling, dialogue systems and information retrieval systems. The training of an automatic speech recognition system is usually straight forward given a large annotated speech corpus for acoustic modeling, a phonetic lexicon, and a text corpus for the training of a language model. However, in some use cases these resources are not available. In this work, we discuss the training of a Swiss German speech recognition system. The only resources that are available is a small size audio corpus, containing the utterances of highly dialectical Swiss German speakers, annotated with a standard German transcription. The desired output of the speech recognizer is again standard German, since there is no other official or standardized way to write Swiss German. We explore strategies to cope with the mismatch between the dialectal pronunciation and the standard German annotation. A Swiss German speech recognizer is trained by adapting a standard German model, based on a Swiss German grapheme-to-phoneme conversion model, which was learned in a data-driven manner. Also, Swiss German speech recognition systems are created, with the pronunciation based on graphemes, standard German pronunciation and with a data-driven Swiss German pronunciation model. The results of the experiments are promising for this challenging task.

Simulating ASR errors for training SLU systems

Edwin Simonnet, Sahar Ghannay, Nathalie Camelin and Yannick Estève

This paper presents an approach to simulate automatic speech recognition (ASR) errors from manual transcriptions and describes how it can be used to improve the performance of spoken language understanding (SLU) systems. In particular, we point out that this noising process is very usefull to obtain a more robust SLU system to ASR errors in case of insufficient training data or more if ASR transcriptions are not available during the training of the SLU model. The proposed method is based on the use of both acoustic and linguistic word embeddings in order to define a similarity measure between words dedicated to predict ASR confusions. Actually, we assume that words acoustically and linguistically close are the ones confused by an ASR system. By using this similarity measure in order to randomly substitute correct words by potentially confusing words in manual annotations used to train CRF- or neural- based SLU systems, we augment the training corpus with these new noisy data. Experiments were carried on the French MEDIA corpus focusing on hotel reservation. They show that this approach significantly improves SLU system performance with a relative reduction of 21.2% of concept/value error rate (CVER), particularly when the SLU system is based on a neural approach (reduction of 22.4% of CVER). A comparison to a naive noising approach shows that the proposed noising approach is particularly relevant.

Evaluation of Feature-Space Speaker Adaptation for End-to-End Acoustic Models

Natalia Tomashenko and Yannick Estève

This paper investigates speaker adaptation techniques for bidirectional long short term memory (BLSTM) recurrent neural network based acoustic models (AMs) trained with the connectionist temporal classification (CTC) objective function. BLSTM-CTC AMs play an important role in end-to-end automatic speech recognition systems. However, there is a lack of research in speaker adaptation algorithms for these models. We explore three different feature-space adaptation approaches for CTC AMs: feature-space maximum linear regression, i-vector based adaptation, and maximum a posteriori adaptation using GMM-derived features. Experimental results on the TED-LIUM corpus demonstrate that speaker adaptation, applied in combination with data augmentation techniques, provides, in an unsupervised adaptation mode, for different test sets, up to 11–20% of relative word error rate reduction over the baseline model built on the raw filter-bank features. In addition, the adaptation behavior is compared for BLSTM-CTC AMs and time-delay neural network AMs trained with the cross-entropy criterion.

Creating New Language and Voice Components for the Updated MaryTTS Text-to-Speech Synthesis Platform

Ingmar Steiner and Sébastien Le Maguer

We present a new workflow to create components for the MaryTTS TTS platform, which is popular with researchers and developers, extending it to support new languages and custom synthetic voices. This workflow replaces the previous toolkit with an efficient, flexible process that leverages modern build automation and cloud-hosted infrastructure. Moreover, it is compatible with the updated MaryTTS architecture, enabling new features and state-of-the-art paradigms such as synthesis based on DNN. Like MaryTTS itself, the new tools are FOSS, and promote the use of open data.

Speech Rate Calculations with Short Utterances: A Study from a Speech-to-Speech, Machine Translation Mediated Map Task

Akira Hayakawa, Carl Vogel, Saturnino Luz and Nick Campbell

The motivation for this paper is to present a way to verify if an utterance within a corpus is pronounced at a fast or slow pace. An alternative method to the well-known Word-Per-Minute (wpm) method for cases where this approach is not applicable. For long segmentations, such as the full introduction section of a speech or presentation, the measurement of wpm is a viable option. For short comparisons of the same single word or multiple syllables, Syllables-Per-Second (sps) is also a viable option. However, when there are multiple short utterances that are frequent in task oriented dialogues or natural free flowing conversation, such as those of the direct Human-to-Human dialogues of the HCRC Map Task corpus or the computer mediated inter-lingual dialogues of the ILMT-s2s corpus, it becomes difficult to obtain a meaningful value for the utterance speech rate. In this paper we explain the method used to provide a alternative speech rate value to the utterance of the ILMT-s2s corpus and the HCRC Map Task corpus.

Session P44 - Summarisation

10th May 2018, 16:50

Chair person: **Senja Pollak** Poster Session

Beyond Generic Summarization: A Multi-faceted Hierarchical Summarization Corpus of Large Heterogeneous Data

Christopher Tauchmann, Thomas Arnold, Andreas Hanselowski, Christian M. Meyer and Margot Mieskes

Automatic summarization has so far focused on datasets of ten to twenty rather short documents, typically news articles. But automatic systems could in theory analyze hundreds of documents from a wide range of sources and provide an overview to the interested reader. Such a summary would ideally present the most general issues of a given topic and allow for more in-depth information on specific aspects within said topic. In this paper, we present a new approach for creating hierarchical summarization corpora from large, heterogeneous document collections. We first extract relevant content using crowdsourcing and then ask trained annotators to order the relevant information hierarchically. This yields tree structures covering the specific facets discussed in a document collection. Our resulting corpus is freely available and can be used to develop and evaluate hierarchical summarization systems.

A New Annotated Portuguese/Spanish Corpus for the Multi-Sentence Compression Task

Elvys Linhares Pontes, Juan-Manuel Torres-Moreno, Stéphane Huet and Andréa carneiro Linhares

Multi-sentence compression aims to generate a short and informative compression from several source sentences that deal with the same topic. In this work, we present a new corpus for the Multi-Sentence Compression (MSC) task in Portuguese and Spanish. We also provide on this corpus a comparison of two state-of-the-art MSC systems.

Live Blog Corpus for Summarization

Avinesh PVS, Maxime Peyrard and Christian M. Meyer

Live blogs are an increasingly popular news format to cover breaking news and live events in online journalism. Online news websites around the world are using this medium to give their readers a minute by minute update on an event. Good summaries enhance the value of the live blogs for a reader but are often not available. In this paper, we study a way of collecting corpora for automatic live blog summarization. In an empirical evaluation using well-known state-of-the-art summarization systems, we show that live blogs corpus poses new challenges in the field of summarization. We make our tools publicly available to reconstruct the corpus to encourage the research community and replicate our results.

TSix: A Human-involved-creation Dataset for Tweet Summarization

Minh-Tien Nguyen, Dac Viet Lai, Huy-Tien Nguyen and Minh-Le Nguyen

We present a new dataset for tweet summarization. The dataset includes six events collected from Twitter from October 10 to November 9, 2016. Our dataset features two prominent properties. Firstly, human-annotated gold-standard references allow to correctly evaluate extractive summarization methods. Secondly, tweets are assigned into sub-topics divided by consecutive days, which facilitate incremental tweet stream summarization methods. To reveal the potential usefulness of our dataset, we compare several well-known summarization methods. Experimental results indicate that among extractive approaches, hybrid term frequency – document term frequency obtains competitive results in term of ROUGE-scores. The analysis also shows that polarity is an implicit factor of tweets in our dataset, suggesting that it can be exploited as a component besides tweet content quality in the summarization process.

A Workbench for Rapid Generation of Cross-Lingual Summaries

Nisarg Jhaveri, Manish Gupta and Vasudeva Varma

The need for cross-lingual information access is more than ever with the easy accessibility to the Internet, especially in vastly multilingual societies like India. Cross-lingual summarization can help minimize human effort needed for achieving publishable articles in multiple languages, while making the most important information available in target language in the form of summaries. We describe a flexible, web-based tool for human editing of cross-lingual summaries to rapidly generate publishable summaries in a number of Indian Languages for news articles originally published in English, and simultaneously collect detailed logs about the process, at both article and sentence level. Similar to translation post-editing logs, such logs can be used to evaluate the automated cross-lingual summaries, in terms of effort needed to make them publishable. The generated summaries along with the logs can be used to train and improve the automatic system over time.

Annotation and Analysis of Extractive Summaries for the Kyutech Corpus

Takashi Yamamura and Kazutaka Shimada

Summarization of multi-party conversation is one of the important tasks in natural language processing. For conversation summarization tasks, corpora have an important role to analyze characteristics of conversations and to construct a method for summary generation. We are developing a freely available Japanese conversation corpus for a decision-making task. We call it the Kyutech corpus. The current version of the Kyutech corpus contains topic tags of each utterance and reference summaries of each conversation. In this paper, we explain an annotation task of extractive summaries. In the annotation task, we annotate an importance tag for each utterance and link utterances with sentences in reference summaries that already exist in the Kyutech corpus. By using the annotated extractive summaries, we can evaluate extractive summarization methods on the Kyutech corpus. In the experiment, we compare some methods based on machine learning techniques with some features.

A Repository of Corpora for Summarization

Franck Dernoncourt, Mohammad Ghassemi and Walter Chang

Summarization corpora are numerous but fragmented, making it challenging for researchers to efficiently pinpoint corpora most suited to a given summarization task. In this paper, we introduce a repository containing corpora available to train and evaluate automatic summarization systems. We also present an overview of the main corpora with respect to the different summarization tasks, and identify various corpus parameters that researchers may want to consider when choosing a corpus. Lastly, as the recent successes of artificial neural networks for summarization have renewed the interest in creating large-scale corpora for summarization, we survey which corpora are used in neural network research studies. We come to the conclusion that more large-scale corpora for summarization are needed. Furthermore, each corpus is organized differently, which makes it time-consuming for researchers to experiment a new summarization algorithm on many corpora, and as a result studies typically use one or very few corpora. Agreeing on a data standard for summarization corpora would be beneficial to the field.

Auto-hMDS: Automatic Construction of a Large Heterogeneous Multilingual Multi-Document Summarization Corpus

Markus Zopf

Automatic text summarization is a challenging natural language processing (NLP) task which has been researched for several decades. The available datasets for multi-document summarization (MDS) are, however, rather small and usually focused on the newswire genre. Nowadays, machine learning methods are applied to more and more NLP problems such as machine translation, question answering, and single-document summarization. Modern machine learning methods such as neural networks require large training datasets which are available for the three tasks but not yet for MDS. This lack of training data limits the development of machine learning methods for MDS. In this work, we automatically generate a large heterogeneous multilingual multi-document summarization corpus. The key idea is to use Wikipedia articles as summaries and to automatically search for appropriate source documents. We created a corpus with 7,316 topics in English and German, which has variing summary lengths and variing number of source documents. More information about the corpus can be found at the corpus GitHub page at https://github.com/AIPHES/auto-hMDS.

PyrEval: An Automated Method for Summary Content Analysis

Yanjun Gao, Andrew Warner and Rebecca Passonneau

Pyramid method is an existing content analysis approach in automatic summarization evaluation for manual construction of a pyramid content model from reference summaries, and manual scoring of the target summaries with the pyramid model. PyrEval assesses the content of automatic summarization by automating the manual pyramid method. PyrEval uses low-dimension distributional semantics to represent phrase meanings,

and a new algorithm, EDUA (Emergent Discoveries of Units of Attractions), for solving set packing problem in construction of content model from vectorized phrases. Because the vectors are pretrained, and EDUA is an efficient greedy algorithm, PyrEval can replace manual pyramid with no retraining, and is very efficient. Moreover, PyrEval has been tested on many datasets derived from humans and machine translated summaries and shown good performance on both.

Session P45 - Textual Entailment and Paraphrasing

10th May 2018, 16:50

Chair person: **Stefan Evert** Poster Session

Mapping Texts to Scripts: An Entailment Study

Simon Ostermann, Hannah Seitz, Stefan Thater and Manfred Pinkal

Commonsense knowledge as provided by scripts is crucially relevant for text understanding systems, providing a basis for commonsense inference. This paper considers a relevant subtask of script-based text understanding, the task of mapping event mentions in a text to script events. We focus on script representations where events are associated with paraphrase sets, i.e. sets of crowdsourced event descriptions. We provide a detailed annotation of event mention/description pairs with textual entailment types. We demonstrate that representing events in terms of paraphrase sets can massively improve the performance of text-to-script mapping systems. However, for a residual substantial fraction of cases, deeper inference is still required.

Semantic Equivalence Detection: Are Interrogatives Harder than Declaratives?

João Rodrigues, Chakaveh Saedi, António Branco and João Silva

Duplicate Question Detection (DQD) is a Natural Language Processing task under active research, with applications to fields like Community Question Answering and Information Retrieval. While DQD falls under the umbrella of Semantic Text Similarity (STS), these are often not seen as similar tasks of semantic equivalence detection, with STS being implicitly understood as concerning only declarative sentences. Nevertheless, approaches to STS have been applied to DQD and paraphrase detection, that is to interrogatives and declaratives, alike. We present a study that seeks to assess, under conditions of comparability, the possible different performance of state-of-the-art approaches to STS over different types of textual segments, including most notably declaratives and interrogatives. This paper contributes to a better understanding of current mainstream methods for semantic equivalence detection, and to a better appreciation of the different results reported in the literature when these are obtained from different data sets with different types of textual segments. Importantly, it contributes also with results concerning how data sets containing textual segments of a certain type can be used to leverage the performance of resolvers for segments of other types.

CEFR-based Lexical Simplification Dataset

Satoru Uchida, Shohei Takada and Yuki Arase

This study creates a language dataset for lexical simplification based on Common European Framework of References for Languages (CEFR) levels (CEFR-LS). Lexical simplification has continued to be one of the important tasks for language learning and education. There are several language resources for lexical simplification that are available for generating rules and creating simplifiers using machine learning. However, these resources are not tailored to language education with word levels and lists of candidates tending to be subjective. Different from these, the present study constructs a CEFR-LS whose target and candidate words are assigned CEFR levels using CEFR-J wordlists and English Vocabulary Profile, and candidates are selected using an online thesaurus. Since CEFR is widely used around the world, using CEFR levels makes it possible to apply a simplification method based on our dataset to language education directly. CEFR-LS currently includes 406 targets and 4912 candidates. To evaluate the validity of CEFR-LS for machine learning, two basic models are employed for selecting candidates and the results are presented as a reference for future users of the dataset.

Session O29 - Language Resource Infrastructures

11th May 2018, 09:45

Chair person: **Takenobu Tokunaga** Oral Session

CLARIN: Towards FAIR and Responsible Data Science Using Language Resources

Franciska De Jong, Bente Maegaard, Koenraad De Smedt, Darja Fišer and Dieter Van Uytvanck

CLARIN is a European Research Infrastructure providing access to language resources and technologies for researchers in the humanities and social sciences. It supports the study of language data in general and aims to increase the potential for comparative research of cultural and societal phenomena across the boundaries of languages. This paper outlines the CLARIN vision and strategy, and it explains how the design and implementation of CLARIN are compliant with the FAIR principles: findability, accessibility, interoperability and re- usability of data. The paper also explains

the approach of CLARIN towards the enabling of responsible data science. Attention is paid to (i) the development of measures for increasing the transparency and explainability of the results from applying CLARIN technologies, in particular in the context of multidisciplinary research, and (ii) stimulating the uptake of its resources, tools and services by the various communities of use, all in accordance with the principles for Open Science.

From 'Solved Problems' to New Challenges: A Report on LDC Activities

Christopher Cieri, Mark Liberman, Stephanie Strassel, Denise DiPersio, Jonathan Wright and Andrea Mazzucchi

This paper reports on the activities of the Linguistic Data Consortium, the next in a sequence of such data center reports included in each LREC meeting. This report begins by sketching the changing demands for Language Resources driven by the spread of Human Language Technologies throughout the market. One result of the successful deployment of HLT enabled applications is increased demand in ever more languages. This in turn places pressure on data centers to collaborate and form global networks in order to meet the demand for LRs of increasing complexity and linguistic diversity. The report next summarizes the over 100 Language Resources released since the last report, many of which have been contributed by research groups around the world. It also covers advances in Consortium infrastructure that assure the integrity of published data sets and support future collection and annotation. Finally, it discusses recent and current LR creation activities that lead to new LR publications followed by data related research activities particularly in clinical applications.

New directions in ELRA activities

Valérie Mapelli, Victoria Arranz, Hélène Mazo, Pawel Kamocki and Vladimir Popescu

Beyond the generic activities (cataloguing, producing, distribution of Language Resources, dissemination of information, etc.) that make the overall ELRA mission an indispensable middle-man in the field of Language Resources (LRs), new directions of work are now being undertaken so as to answer the needs of this ever-moving community. This impacts the structure and the operating model of the association per se with the creation of a new technical committee dealing with Less-Resourced Languages and the modification of the ELRA membership policy. It also intrinsically impacts the axes of work at all steps of activities: offering new tools for sharing LRs and related information, adapting to new legal requirements, producing and offering field-specific data. This paper addresses these new directions and describes ELRA (and its operational body ELDA) regular activities updates. Future activities are also reported in the last part of the article. They consist in ongoing projects like the ELRC initiative, the start of another CEF-funded project, the European Language Resource Infrastructure (ELRI), the updating of the Review of existing Language Resources for languages of France, and the continuation of the ELRA Catalogue development.

A Framework for Multi-Language Service Design with the Language Grid

Donghui Lin, Yohei Murakami and Toru Ishida

To collect and share language resources like machine translators and dictionaries, we developed the Language Grid in 2006, a service-oriented language infrastructure on the Internet. Although we have put a lot of effort into improving the service grid technologies and collecting language services, international NPO/NGOs are struggling with the design and development of tools and systems for supporting multi-language communication in the real world by utilizing available language services. This paper proposes a framework for service design with the Language Grid by bridging the gap between language service infrastructures and multi-language systems. The proposed framework is implemented as a toolkit, Multilingual Studio, which is open to allow the users to design and develop multilingual communication services and tools in the real world.

Language Technology for Multilingual Europe: An Analysis of a Large-Scale Survey regarding Challenges, Demands, Gaps and Needs

Georg Rehm and Stefanie Hegele

We present the analysis of a large-scale survey titled "Language Technology for Multilingual Europe", conducted between May and June 2017. A total of 634 participants in 52 countries responded to the survey. Its main purpose was to collect input, feedback and ideas from the European Language Technology research and innovation community in order to assess the most prominent research areas, projects and applications, but, more importantly to identify the biggest challenges, obstacles and gaps Europe is currently facing with regard to its multilingual setup and technological solutions. Participants were encouraged to share concrete suggestions and recommendations on how present challenges can be turned into opportunities in the context of a potential long-term, large-scale, Europe-wide research, development and innovation funding programme, currently titled Human Language Project.

Session O30 - Digital Humanities & Text Analytics

11th May 2018, 09:45

Chair person: **Thierry Declerck** Oral Session

Annotating High-Level Structures of Short Stories and Personal Anecdotes

Boyang Li, Beth Cardier, Tong Wang and Florian Metze

Stories are a vital form of communication in human culture; they are employed daily to persuade, to elicit sympathy, or to convey a message. Computational understanding of human narratives, especially high-level narrative structures, remain limited to date. Multiple literary theories for narrative structures exist, but operationalization of the theories has remained a challenge. We developed an annotation scheme by consolidating and extending existing narratological theories, including Labov and Waletsky's (1967) functional categorization scheme and Freytag's (1863) pyramid of dramatic tension, and present 360 annotated short stories collected from online sources. In the future, this research will support an approach that enables systems to intelligently sustain complex communications with humans.

Discovering the Language of Wine Reviews: A Text Mining Account

Els Lefever, Iris Hendrickx, Ilja Croijmans, Antal Van den Bosch and Asifa Majid

It is widely held that smells and flavors are impossible to put into words. In this paper we test this claim by seeking predictive patterns in wine reviews, which ostensibly aim to provide guides to perceptual content. Wine reviews have previously been critiqued as random and meaningless. We collected an English corpus of wine reviews with their structured metadata, and applied machine learning techniques to automatically predict the wine's color, grape variety, and country of origin. To train the three supervised classifiers, three different information sources were incorporated: lexical bag-of-words features, domain-specific terminology features, and semantic word embedding features. In addition, using regression analysis we investigated basic review properties, i.e., review length, average word length, and their relationship to the scalar values of price and review score. Our results show that wine experts do share a common vocabulary to describe wines and they use this in a consistent way, which makes it possible to automatically predict wine characteristics based on the review text alone. This means that odors and flavors may be more expressible in language than typically acknowledged.

Toward An Epic Epigraph Graph

Francis Bond and Graham Matthews

We present a database of epigraphs collected with the goal of revealing literary influence as a set of connections between authors over time. We have collected epigraphs from over 12,000 literary works and are in the process of identifying their provenance. The database is released under an open license.

Delta vs. N-Gram Tracing: Evaluating the Robustness of Authorship Attribution Methods

Thomas Proisl, Stefan Evert, Fotis Jannidis, Christof Schöch, Leonard Konle and Steffen Pielström

Delta measures are a well-established and popular family of authorship attribution methods, especially for literary texts. N-gram tracing is a novel method for authorship attribution designed for very short texts, which has its roots in forensic linguistics. We evaluate the performance of both methods in a series of experiments on English, French and German literary texts, in order to investigate the relationship between authorship attribution accuracy and text length as well as the composition of the comparison corpus. Our results show that, at least in our setting, both methods require relatively long texts and are furthermore highly sensitive to the choice of authors and texts in the comparison corpus.

An Attribution Relations Corpus for Political News

Edward Newell, Drew Margolin and Derek Ruths

An attribution occurs when an author quotes, paraphrases, or describes the statements and private states of a third party. Journalists use attribution to report statements and attitudes of public figures, organizations, and ordinary individuals. Properly recognizing attributions in context is an essential aspect of natural language understanding and implicated in many NLP tasks, but current resources are limited in size and completeness. We introduce the Political News Attribution Relations Corpus 2016 (PolNeAR)—the largest, most complete attribution relations corpus to date. This dataset greatly increases the volume of high-quality attribution annotations, addresses shortcomings of existing resources, and expands the diversity of publishers sourced. PolNeAR is built on news articles covering the political candidates during the year leading up to US Presidential Election in November of 2016. The dataset will support the creation of sophisticated end-to-end solutions for attribution extraction and invite interdisciplinary collaboration between the NLP, communications, political science, and journalism communities. Along with the dataset we contribute revised guidelines aimed at improving clarity and consistency in the annotation task, and an

annotation interface specially adapted to the task, for reproduction or extension of this work

Session O31 - Crowdsourcing & Collaborative Resource Construction

11th May 2018, 09:45

Chair person: **Steve Cassidy** Oral Session

Face2Text: Collecting an Annotated Image Description Corpus for the Generation of Rich Face Descriptions

Albert Gatt, Marc Tanti, Adrian Muscat, Patrizia Paggio, Reuben A Farrugia, Claudia Borg, Kenneth Camilleri, Mike Rosner and Lonneke Van der Plas

The past few years have witnessed renewed interest in NLP tasks at the interface between vision and language. One intensively-studied problem is that of automatically generating text from images. In this paper, we extend this problem to the more specific domain of face description. Unlike scene descriptions, face descriptions are more fine-grained and rely on attributes extracted from the image, rather than objects and relations. Given that no data exists for this task, we present an ongoing crowdsourcing study to collect a corpus of descriptions of face images taken 'in the wild'. To gain a better understanding of the variation we find in face description and the possible issues that this may raise, we also conducted an annotation study on a subset of the corpus. Primarily, we found descriptions to refer to a mixture of attributes, not only physical, but also emotional and inferential, which is bound to create further challenges for current image-to-text methods.

Adapting Serious Game for Fallacious Argumentation to German: Pitfalls, Insights, and Best Practices

Ivan Habernal, Patrick Pauli and Iryna Gurevych

As argumentation about controversies is culture- and language-dependent, porting a serious game that deals with daily argumentation to another language requires substantial adaptation. This article presents a study of deploying Argotario (serious game for learning argumentation fallacies) in the German context. We examine all steps that are necessary to end up with a successful serious game platform, such as topic selection, initial data creation, or effective campaigns. Moreover, we analyze users' behavior and in-game created data in order to assess the dissemination strategies and qualitative aspects of the resulting corpus. We also report on classification experiments based on neural networks and feature-based models.

Crowdsourcing Regional Variation Data and Automatic Geolocalisation of Speakers of European French

Jean-Philippe Goldman, Yves Scherrer, Julie Glikman, Mathieu Avanzi, Christophe Benzitoun and Philippe Boula de Mareüil

We present the crowdsourcing platform Donnez Votre Français à la Science (DFS, or "Give your French to Science"), which aims to collect linguistic data and document language use, with a special focus on regional variation in European French. The activities not only gather data that is useful for scientific studies, but they also provide feedback to the general public; this is important in order to reward participants, to encourage them to follow future surveys, and to foster interaction with the scientific community. The two main activities described here are 1) a linguistic survey on lexical variation with immediate feedback and 2) a speaker geolocalisation system; i.e., a quiz that guesses the linguistic origin of the participant by comparing their answers with previously gathered linguistic data. For the geolocalisation activity, we set up a simulation framework to optimise predictions. Three classification algorithms are compared: the first one uses clustering and shibboleth detection, whereas the other two rely on feature elimination techniques with Support Vector Machines and Maximum Entropy models as underlying base classifiers. The best-performing system uses a selection of 17 questions and reaches a localisation accuracy of 66%, extending the prediction from the one-best area (one among 109 base areas) to its first-order and second-order neighbouring areas.

Improving Machine Translation of Educational Content via Crowdsourcing

Maximiliana Behnke, Antonio Valerio Miceli Barone, Rico Sennrich, Vilelmini Sosoni, Thanasis Naskos, Eirini Takoulidou, Maria Stasimioti, Menno Van Zaanen, Sheila Castilho, Federico Gaspari, Panayota Georgakopoulou, Valia Kordoni, Markus Egg and Katia Lida Kermanidis

The limited availability of in-domain training data is a major issue in the training of application-specific neural machine translation models. Professional outsourcing of bilingual data collections is costly and often not feasible. In this paper we analyze the influence of using crowdsourcing as a scalable way to obtain translations of target in-domain data having in mind that the translations can be of a lower quality. We apply crowdsourcing with carefully designed quality controls to create parallel corpora for the educational domain by collecting translations of texts from MOOCs from English to eleven languages, which we then use to fine-tune neural machine translation models previously trained on general-domain data. The results from our research

indicate that crowdsourced data collected with proper quality controls consistently yields performance gains over general-domain baseline systems, and systems fine-tuned with pre-existing in-domain corpora.

Grounding Gradable Adjectives through Crowdsourcing

Rebecca Sharp, Mithun Paul, Ajay Nagesh, Dane Bell and Mihai Surdeanu

In order to build technology that has the ability to answer questions relevant to national and global security, e.g., on food insecurity in certain parts of the world, one has to implement machine reading technology that extracts causal mechanisms from texts. Unfortunately, many of these texts describe these interactions using vague, high-level language. One particular example is the use of gradable adjectives, i.e., adjectives that can take a range of magnitudes such as small or slight. Here we propose a method for estimating specific concrete groundings for a set of such gradable adjectives. We use crowdsourcing to gather human language intuitions about the impact of each adjective, then fit a linear mixed effects model to this data. The resulting model is able to estimate the impact of novel instances of these adjectives found in text. We evaluate our model in terms of its ability to generalize to unseen data and find that it has a predictive R2 of 0.632 in general, and 0.677 on a subset of high-frequency adjectives.

Session O32 - Less-Resourced Languages Speech & Multimodal Corpora

11th May 2018, 09:45

Chair person: **Shyam Agrawal** Oral Session

Evaluation Phonemic Transcription of Low-Resource Tonal Languages for Language Documentation

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird and Alexis MICHAUD

Transcribing speech is an important part of language documentation, yet speech recognition technology has not been widely harnessed to aid linguists. We explore the use of a neural network architecture with the connectionist temporal classification loss function for phonemic and tonal transcription in a language documentation setting. In this framework, we explore jointly modelling phonemes and tones versus modelling them separately, and assess the importance of pitch information versus phonemic context for tonal prediction. Experiments on two tonal languages, Yongning Na and Eastern Chatino, show the changes in recognition performance as training data is scaled from 10 minutes up to 50 minutes for Chatino, and up to 224 minutes for Na. We discuss the findings from incorporating this technology into the linguistic workflow for documenting Yongning Na, which show the method's promise in improving efficiency, minimizing typographical errors, and maintaining the transcription's faithfulness to the acoustic signal, while highlighting phonetic and phonemic facts for linguistic consideration.

A Very Low Resource Language Speech Corpus for Computational Language Documentation Experiments

Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noel Kouarata, Lori Lamel, Hélène Maynard, Markus Mueller, Annie Rialland, Sebastian Stueker, François Yvon and Marcely Zanon Boito

Most speech and language technologies are trained with massive amounts of speech and text information. However, most of the world languages do not have such resources and some even lack a stable orthography. Building systems under these almost zero resource conditions is not only promising for speech technology but also for computational language documentation. The goal of computational language documentation is to help field linguists to (semi-)automatically analyze and annotate audio recordings of endangered, unwritten languages. Example tasks are automatic phoneme discovery or lexicon discovery from the speech signal. This paper presents a speech corpus collected during a realistic language documentation process. It is made up of 5k speech utterances in Mboshi (Bantu C25) aligned to French text translations. Speech transcriptions are also made available: they correspond to a non-standard graphemic form close to the language phonology. We detail how the data was collected, cleaned and processed and we illustrate its use through a zero-resource task: spoken term discovery. The dataset is made available to the community for reproducible computational language documentation experiments and their evaluation.

Chahta Anumpa: A multimodal corpus of the Choctaw Language

Jacqueline Brixey, Eli Pincus and Ron Artstein

This paper presents a general use corpus for the Native American indigenous language Choctaw. The corpus contains audio, video, and text resources, with many texts also translated in English. The Oklahoma Choctaw and the Mississippi Choctaw variants of the language are represented in the corpus. The data set provides documentation support for the threatened language, and allows researchers and language teachers access to a diverse collection of resources.

BULBasaa: A Bilingual Basaa-French Speech Corpus for the Evaluation of Language Documentation Tools

Fatima Hamlaoui, Emmanuel-Moselly Makasso, Markus Müller, Jonas Engelmann, Gilles Adda, Alex Waibel and Sebastian Stüker

Basaa is one of the three Bantu languages of BULB (Breaking the Unwritten Language Barrier), a project whose aim is to provide NLP-based tools to support linguists in documenting under-resourced and unwritten languages. To develop technologies such as automatic phone transcription or machine translation, a massive amount of speech data is needed. Approximately 50 hours of Basaa speech were thus collected and then carefully re-spoken and orally translated into French in a controlled environment by a few bilingual speakers. For a subset of approx. 10 hours of the corpus, each utterance was additionally phonetically transcribed to establish a golden standard for the output of our NLP tools. The experiments described in this paper are meant to provide an automatic phonetic transcription using a set of derived phone-like units. As every language features a specific set of idiosyncrasies, automating the process of phonetic unit discovery in its entirety is a challenging task. Within BULB, we envision a workflow where linguists are able to refine the set of automatically discovered units and the system is then able to re-iterate on the data, providing a better approximation of the actual phone set.

Researching Less-Resourced Languages – the DigiSami Corpus

Kristiina Jokinen

Increased use of digital devices and data repositories has enabled a digital revolution in data collection and language research, and has also led to important activities supporting speech and language technology research for less-resourced languages. This paper describes the DigiSami project and its research results, focussing on spoken corpus collection and speech technology for the Fenno-Ugric language North Sami. The paper also discusses multifaceted questions on ethics and privacy related to data collection for less-resourced languages and indigenous communities.

Session P46 - Dialects

11th May 2018, 09:45

Chair person: **Claudia Soria** Poster Session

Creating dialect sub-corpora by clustering: a case in Japanese for an adaptive method

Yo Sato and Kevin Heffernan

We propose a pipeline through which to derive clusters of dialects, given a mixed corpus composed of different dialects, when their standard counterpart is sufficiently resourced. The test case is Japanese, where the written standard language is sufficiently equipped with adequate resources. Our method starts by detecting non-standard contents rst, and then clusters what is deemed dialectal. We report the results on the clustering of mixed Twitter corpus into four dialects (Kansai, Tohoku, Chugoku and Kyushu).

A Fast and Flexible Webinterface for Dialect Research in the Low Countries

Roeland Van Hout, Nicoline Van der Sijs, Erwin Komen and Henk Van den Heuvel

This paper describes the development of webportals with search applications built in order to make the data from the 33 volumes of the Dictionary of the Brabantic dialects (1967-2005) and the 39 volumes of the Dictionary of the Limburgian dialects (1983-2008) accessible and retrievable for both the research community and the general audience. Part of the data was available in a digital format, a larger part only in print. The printed data was semi-automatically converted from paper to structured text (database). This process allowed for streamlining information, applying (semi-)automatic data checks and manually correcting the input. Next, the resulting database was the backbone of a webportal for faceted search requests on the full collection, including filtering and splitting the results on metadata. The design and implementation of the webportals, called e-WBD and e-WLD, are being defined in more detail. The URLs of the portals are: http://e-wbd.nl/ and http://www.e-wld.nl/.

Arabic Dialect Identification in the Context of Bivalency and Code-Switching

Mahmoud El-Haj, Paul Rayson and Mariam Aboelezz

In this paper we use a novel approach towards Arabic dialect identification using language bivalency and written code-switching. Bivalency between languages or dialects is where a word or element is treated by language users as having a fundamentally similar semantic content in more than one language or dialect. Arabic dialect identification in writing is a difficult task even for humans due to the fact that words are used interchangeably between dialects. The task of automatically identifying dialect is harder and classifiers trained using only n-grams will perform poorly when tested on unseen data. Such approaches require significant amounts of annotated training data which is costly and time consuming to produce. Currently available Arabic dialect datasets do not exceed a few hundred thousand sentences, thus we need to extract features other than word and character n-grams. In our work we present experimental results from automatically identifying dialects from the four main Arabic dialect regions (Egypt, North Africa, Gulf and Levant)

in addition to Standard Arabic. We extend previous work by incorporating additional grammatical and stylistic features and define a subtractive bivalency profiling approach to address issues of bivalent words across the examined Arabic dialects. The results show that our new methods classification accuracy can reach more than 76% and score well (66%) when tested on completely unseen data.

Unified Guidelines and Resources for Arabic Dialect Orthography

Nizar Habash, Fadhl Eryani, Salam Khalifa, Owen Rambow, Dana Abdulrahim, Alexander Erdmann, Reem Faraj, Wajdi Zaghouani, Houda Bouamor, Nasser Zalmout, Sara Hassan, Faisal Al shargi, Sakhar Alkhereyf, Basma Abdulkareem, Ramy Eskander, Mohammad Salameh and Hind Saddiki

We present a unified set of guidelines and resources for conventional orthography of dialectal Arabic. While Standard Arabic has well defined orthographic standards, none of the Arabic dialects do today. Previous efforts on conventionalizing the dialectal orthography have focused on specific dialects and made often ad hoc decisions. In this work, we present a common set of guidelines and meta-guidelines and apply them to 28 Arab city dialects from Rabat to Muscat. These guidelines and their connected resources are being used by three large Arabic dialect processing projects in three universities.

Automatic Identification of Maghreb Dialects Using a Dictionary-Based Approach

Houda SAADANE, Hosni Seffih, Christian Fluhr, Khalid Choukri and Nasredine SEMMAR

Automatic identification of Arabic dialects in a text is a difficult task, especially for Maghreb languages and when they are written in Arabic or Latin characters (Arabizi). These texts are characterized by the use of code-switching between the Modern Standard Arabic (MSA) and the Arabic Dialect (AD) in the texts written in Arabic, or between Arabizi and foreign languages for those written in Latin. This paper presents the specific resources and tools we have developed for this purpose, with a focus on the transliteration of Arabizi into Arabic (using the dedicated tools for Arabic dialects). A dictionary-based approach to detect the dialectal origin of a text is described, it exhibits satisfactory results.

Shami: A Corpus of Levantine Arabic Dialects

Chatrine Qwaider, Motaz Saad, Stergios Chatzikyriakidis and Simon Dobnik

Modern Standard Arabic (MSA) is the official language used in education and media across the Arab world both in writing and formal speech. However, in daily communication several dialects depending on the country, region as well as other social factors, are used. With the emergence of social media, the dialectal amount of data on the Internet have increased and the NLP tools that support MSA are not well-suited to process this data due to the difference between the dialects and MSA. In this paper, we construct the Shami corpus, the first Levantine Dialect Corpus (SDC) covering data from the four dialects spoken in Palestine, Jordan, Lebanon and Syria. We also describe rules for pre-processing without affecting the meaning so that it is processable by NLP tools. We choose Dialect Identification as the task to evaluate SDC and compare it with two other corpora. In this respect, experiments are conducted using different parameters based on n-gram models and Naive Bayes classifiers. SDC is larger than the existing corpora in terms of size, words and vocabularies. In addition, we use the performance on the Language Identification task to exemplify the similarities and differences in the individual dialects

You Tweet What You Speak: A City-Level Dataset of Arabic Dialects

Muhammad Abdul-Mageed, Hassan Alhuzali and Mohamed Elaraby

Arabic has a wide range of varieties or dialects. Although a number of pioneering works have targeted some Arabic dialects, other dialects remain largely without investigation. A serious bottleneck for studying these dialects is lack of any data that can be exploited in computational models. In this work, we aim to bridge this gap: We present a considerably large dataset of > 1=4 billion tweets representing a wide range of dialects. Our dataset is more nuanced than previously reported work in that it is labeled at the fine-grained level of city. More specifically, the data represent 29 major Arab cities from 10 Arab countries with varying dialects (e.g., Egyptian, Gulf, KSA, Levantine, Yemeni).

Visualizing the "Dictionary of Regionalisms of France" (DRF)

Ada Wan

This paper presents CorpusDRF, an open-source, digitized collection of regionalisms, their parts of speech and recognition rates, published in 'Dictionnaire des Regionalismes de France' (DRF, "Dictionary of Regionalisms of France") (Rezeau, 2001), enabling the visualization and analyses of the largest-scale study of French regionalisms in the 20th century using publicly available data. CorpusDRF was curated and checked manually against the entirety of the printed volume of more than 1000 pages. It contains all the entries in the DRF for which recognition rates in continental France were recorded from the surveys carried out from 1994 to 1996 and from 1999 to 2000. In this

paper, in addition to introducing the corpus, we also offer some exploratory visualizations using an easy-to-use, freely available web application and compare the patterns in our analysis with that by (Goebl, 2005a) and (Goebl, 2007).

DART: A Large Dataset of Dialectal Arabic Tweets

Israa Alsarsour, Esraa Mohamed, Reem Suwaileh and Tamer Elsayed

In this paper, we present a new large manually-annotated multi-dialect dataset of Arabic tweets that is publicly available. The Dialectal ARabic Tweets (DART) dataset has about 25K tweets that are annotated via crowdsourcing and it is well-balanced over five main groups of Arabic dialects: Egyptian, Maghrebi, Levantine, Gulf, and Iraqi. The paper outlines the pipeline of constructing the dataset from crawling tweets that match a list of dialect phrases to annotating the tweets by the crowd. We also touch some challenges that we face during the process. We evaluate the quality of the dataset from two perspectives: the inter-annotator agreement and the accuracy of the final labels. Results show that both measures were substantially high for the Egyptian, Gulf, and Levantine dialect groups, but lower for the Iraqi and Maghrebi dialects, which indicates the difficulty of identifying those two dialects manually and hence automatically.

Classification of Closely Related Sub-dialects of Arabic Using Support-Vector Machines

Samantha Wray

Colloquial dialects of Arabic can be roughly categorized into five groups based n relatedness and geographic location (Egyptian, North African/Maghrebi, Gulf, Iraqi, and Levantine), but given that all dialects utilize much of the same writing system and share overlapping features and vocabulary, dialect identification and text classification is no trivial task. Furthermore, text classification by dialect is often performed at a coarse-grained level into these five groups or a subset thereof, and there is little work on sub-dialectal classification. The current study utilizes an n-gram based SVM to classify on a fine-grained sub-dialectal level, and compares it to methods used in dialect classification such as vocabulary pruning of shared items across dialects. A test case of the dialect Levantine is presented here, and results of 65% accuracy on a four-way classification experiment to sub-dialects of Levantine (Jordanian, Lebanese, Palestinian and Syrian) are presented and discussed. This paper also examines the possibility of leveraging existing mixed-dialectal resources to determine their sub-dialectal makeup by automatic classification.

Session P47 - Document Classification, Text Categorisation (2)

11th May 2018, 09:45

Chair person: **Piotr Pęzik** Poster Session

Page Stream Segmentation with Convolutional Neural Nets Combining Textual and Visual Features

Gregor Wiedemann and Gerhard Heyer

In recent years, (retro-)digitizing paper-based files became a major undertaking for private and public archives as well as an important task in electronic mailroom applications. As a first step, the workflow involves scanning and Optical Character Recognition (OCR) of documents. Preservation of document contexts of single page scans is a major requirement in this context. To facilitate workflows involving very large amounts of paper scans, page stream segmentation (PSS) is the task to automatically separate a stream of scanned images into multi-page documents. In a digitization project together with a German federal archive, we developed a novel approach based on convolutional neural networks (CNN) combining image and text features to achieve optimal document separation results. Evaluation shows that our PSS architecture achieves an accuracy up to 93 % which can be regarded as a new state-of-the-art for this task.

Automating Document Discovery in the Systematic Review Process: How to Use Chaff to Extract Wheat

Christopher Norman, Mariska Leeflang, Pierre Zweigenbaum and Aurélie Névéol

Systematic reviews in e.g. empirical medicine address research questions by comprehensively examining the entire published literature. Conventionally, manual literature surveys decide inclusion in two steps, first based on abstracts and title, then by full text, yet current methods to automate the process make no distinction between gold data from these two stages. In this work we compare the impact different schemes for choosing positive and negative examples from the different screening stages have on the training of automated systems. We train a ranker using logistic regression and evaluate it on a new gold standard dataset for clinical NLP, and on an existing gold standard dataset for drug class efficacy. The classification and ranking achieves an average AUC of 0.803 and 0.768 when relying on gold standard decisions based on title and abstracts of articles, and an AUC of 0.625 and 0.839 when relying on gold standard decisions based on full text. Our results suggest that it makes little difference which screening stage the gold standard decisions are drawn from, and

that the decisions need not be based on the full text. The results further suggest that common-off-the-shelf algorithms can reduce the amount of work required to retrieve relevant literature.

Two Multilingual Corpora Extracted from the Tenders Electronic Daily for Machine Learning and Machine Translation Applications.

Oussama Ahmia, Nicolas Béchet and Pierre-François Marteau

The European "Tenders Electronic Daily" (TED) is a large source of semi-structured and multilingual data that is very valuable to the Natural Language Processing community. This data sets can effectively be used to address complex machine translation, multilingual terminology extraction, text-mining, or to benchmark information retrieval systems. Despite of the services offered by the user-friendliness of the web site that is made available to the public to access the publishing of the EU call for tenders, collecting and managing such kind of data is a great burden and consumes a lot of time and computing resources. This could explain why such a resource is not very (if any) exploited today by computer scientists or engineers in NLP. The aim of this paper is to describe two documented and easy-to-use multilingual corpora (one of them is a parallel corpus), extracted from the TED web source that we will release for the benefit of the NLP community.

Using Adversarial Examples in Natural Language Processing

Petr Bělohlávek, Ondřej Plátek, Zdeněk Žabokrtský and Milan Straka

Machine learning models have been providing promising results in many fields including natural language processing. These models are, nevertheless, prone to adversarial examples. These are artificially constructed examples which evince two main features: they resemble the real training data but they deceive already trained model. This paper investigates the effect of using adversarial examples during the training of recurrent neural networks whose text input is in the form of a sequence of word/character embeddings. The effects are studied on a compilation of eight NLP datasets whose interface was unified for quick experimenting. Based on the experiments and the dataset characteristics, we conclude that using the adversarial examples for NLP tasks that are modeled by recurrent neural networks provides a regularization effect and enables the training of models with greater number of parameters without overfitting. In addition, we discuss which combinations of datasets and model settings might benefit from the adversarial training the most.

Modeling Trolling in Social Media Conversations

Luis Gerardo Mojica de la Vega and Vincent Ng

Social media websites, electronic newspapers and Internet forums allow visitors to leave comments for others to read and interact. This exchange is not free from participants with malicious intentions, who troll others by positing messages that are intended to be provocative, offensive, or menacing. With the goal of facilitating the computational modeling of trolling, we propose a trolling categorization that is novel in that it allows comment-based analysis from both the trolls' and the responders' perspectives, characterizing these two perspectives using four aspects, namely, the troll's intention and his intention disclosure, as well as the responder's interpretation of the troll's intention and her response strategy. Using this categorization, we annotate and release a dataset containing excerpts of Reddit conversations involving suspected trolls and their interactions with other users.

Session P48 - Information Extraction, Information Retrieval, Text Analytics (3)

11th May 2018, 09:45

Chair person: **Patrick Drouin** Poster Session

Automatic Annotation of Semantic Term Types in the Complete ACL Anthology Reference Corpus

Anne-Kathrin Schumann and Héctor Martínez Alonso

In the present paper, we present an automated tagging approach aimed at enhancing a well-known resource, the ACL Anthology Reference Corpus, with semantic class labels for more than 20,000 technical terms that are relevant to the domain of computational linguistics. We use state-of-the-art classification techniques to assign semantic class labels to technical terms extracted from several reference term lists. We also sketch a set of research questions and approaches directed towards the integrated analysis of scientific corpora. To this end, we query the data set resulting from our annotation effort on both the term and the semantic class level level.

Annotated Corpus of Scientific Conference's Homepages for Information Extraction

Piotr Andruszkiewicz and Rafal Hazan

In this paper, we present a new corpus that contains 943 homepages of scientific conferences, 14794 including subpages, with annotations of interesting information: name of a conference, its abbreviation, place, and several important dates; that is, submission, notification, and camera ready dates. The topics of conferences included in the corpus are equally distributed over

five areas: artificial intelligence, natural language processing, computer science, telecommunication, and image processing. The corpus is publicly available. Beside the characteristics of the corpus, we present the results of information extraction from the corpus using SVM and CRF models as we would like this corpus to be considered a reference data set for this type of task.

Improving Unsupervised Keyphrase Extraction using Background Knowledge

Yang Yu and Vincent Ng

Keyphrase is an efficient representation of the main idea of documents. While background knowledge can provide valuable information about documents, they are rarely incorporated in keyphrase extraction methods. In this paper, we propose WikiRank, an unsupervised method for keyphrase extraction based on the background knowledge from Wikipedia. Firstly, we construct a semantic graph for the document. Then we transform the keyphrase extraction problem into an optimization problem on the graph. Finally, we get the optimal keyphrase set to be the output. Our method obtains improvements over other state-of-art models by more than 2% in F1-score.

WikiDragon: A Java Framework For Diachronic Content And Network Analysis Of MediaWikis

Rüdiger Gleim, Alexander Mehler and Sung Y. Song

We introduce WikiDragon, a Java Framework designed to give developers in computational linguistics an intuitive API to build, parse and analyze instances of MediaWikis such as Wikipedia, Wiktionary or WikiSource on their computers. It covers current versions of pages as well as the complete revision history, gives diachronic access to both page source code as well as accurately parsed HTML and supports the diachronic exploration of the page network. WikiDragon is self enclosed and only requires an XML dump of the official Wikimedia Foundation website for import into an embedded database. No additional setup is required. We describe WikiDragon's architecture and evaluate the framework based on the simple English Wikipedia with respect to the accuracy of link extraction, diachronic network analysis and the impact of using different Wikipedia frameworks to text analysis.

Studying Muslim Stereotyping through Microportrait Extraction

Antske Fokkens, Nel Ruigrok, Camiel Beukeboom, Gagestein Sarah and Wouter Van Attveldt

Research from communication science has shown that stereotypical ideas are often reflected in language use. Media coverage of different groups in society influences the perception people have about these groups and even increases distrust and polarization among different groups. Investigating the forms of (especially subtle) stereotyping can raise awareness to journalists and help prevent reinforcing oppositions between groups in society. Conducting large-scale, deep investigations to determine whether we are faced with stereotyping is time-consuming and costly. We propose to tackle this challenges through the means of microportraits: an impression of a target group or individual conveyed in a single text. We introduce the first system implementation for Dutch and show that microportraits allow social scientists to explore various dimensions of stereotyping. We explore the possibilities provided by microportraits by investigating stereotyping of Muslims in the Dutch media. Our (preliminary) results show that microportraits provide more detailed insights into stereotyping compared to more basic models such as word clouds.

Analyzing the Quality of Counseling Conversations: the Tell-Tale Signs of High-quality Counseling

Verónica Pérez-Rosas, Xuetong Sun, Christy Li, Yuchen Wang, Kenneth Resnicow and Rada Mihalcea

Behavioral and mental health are pressing issues worldwide. Counseling is emerging as a core treatment for a variety of mental and behavioral health disorders. Seeking to improve the understanding of counseling practice, researchers have started to explore Natural Language Processing approaches to analyze the nature of counseling interactions by studying aspects such as mirroring, empathy, and reflective listening. A challenging aspect of this task is the lack of psychotherapy corpora. In this paper, we introduce a new dataset of high-quality and low-quality counseling conversations collected from public web sources. We present a detailed description of the dataset collection process, including preprocessing, transcription, and the annotation of two counseling micro-skills: reflective listening and questions. We show that the obtained dataset can be used to build text-based classifiers able to predict the overall quality of a counseling conversation and provide insights into the linguistic differences between low and high quality counseling.

Interpersonal Relationship Labels for the CALLHOME Corpus

Denys Katerenchuk, David Guy Brizan and Andrew Rosenberg

The way we speak to our friends, colleagues, or partners is different in both the explicit context, what we say, and the implicit, how we say it. Understanding these differences is important because it provides additional information that can be used in natural language processing tasks. For example, knowing the relationship between interlocutors can help to narrow the range of

topics and improve automatic speech recognition system results. Unfortunately, the lack of corpora makes exploration of this problem intractable. In this work, we release a set of interpersonal relationship labels between conversation participants for the CALLHOME English corpus. We make the labels freely available for download on our website and hope that this effort can further boost research in this area.

Text Mining for History: first steps on building a large dataset

Suemi Higuchi, Cláudia Freitas, Bruno Cuconato and Alexandre Rademaker

This paper presents the initial efforts towards the creation of a new corpus on the history domain. Motivated by the historians' need to interrogate a vast material - almost 12 million words and more than three hundred thousand sentences - in a non-linear way, our approach privileges deep linguistic analysis on an encyclopedic-style data. In this context, the work presented here focuses on the preparation of the corpus, which is prior to the mining activity: the morphosyntactic annotation and the definition of semantic types for entities and relations relevant to the History domain. Taking advantage of the semantic nature of appositive constructions, we manually analyzed a sample of eleven hundred sentences in order to verify its potential as additional semantic clues to be considered. The results show that we are on the right track.

Building Evaluation Datasets for Cultural Microblog Retrieval

Lorraine Goeuriot, Josiane Mothe, Philippe Mulhem and Eric SanJuan

ECLEF Microblog Cultural Contextualization is an evaluation challenge aiming at providing the research community with datasets to gather, organize and deliver relevant social data related to events generating large number of microblogs and web documents. The evaluation challenges runs every year since 2016. We describe in this paper the resources built for the challenge, that can be used outside of the context of the challenge.

Session P49 - Machine Translation, SpeechToSpeech Translation (2)

11th May 2018, 09:45

Chair person: **Mona Diab** Poster Session

Training and Adapting Multilingual NMT for Less-resourced and Morphologically Rich Languages

Matīss Rikters, Mārcis Pinnis and Rihards Krišlauks

In this paper, we present results of employing multilingual and multi-way neural machine translation approaches for morphologically rich languages, such as Estonian and Russian. We experiment with different NMT architectures that allow achieving state-of-the-art translation quality and compare the multi-way model performance to one-way model performance. We report improvements of up to +3.27 BLEU points over our baseline results, when using a multi-way model trained using the transformer network architecture. We also provide open-source scripts used for shuffling and combining multiple parallel datasets for training of the multilingual systems.

Cross-lingual Terminology Extraction for Translation Quality Estimation

YU Yuan, Yuze Gao, Yue Zhang and Serge Sharoff

We explore ways of identifying terms from monolingual texts and integrate them into investigating the contribution of terminology to translation quality.The researchers proposed a supervised learning method using common statistical measures for termhood and unithood as features to train classifiers for identifying terms in cross-domain and cross-language settings. On its basis, sequences of words from source texts (STs) and target texts (TTs) are aligned naively through a fuzzy matching mechanism for identifying the correctly translated term equivalents in student translations. Correlation analyses further show that normalized term occurrences in translations have weak linear relationship with translation quality in term of usefulness/transfer, terminology/style, idiomatic writing and target mechanics and near- and above-strong relationship with the overall translation quality. This method has demonstrated some reliability in automatically identifying terms in human translations. However, drawbacks in handling low frequency terms and term variations shall be dealt in the future.

Machine Translation of Low-Resource Spoken Dialects: Strategies for Normalizing Swiss German

Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat and Michael Baeriswyl

The goal of this work is to design a machine translation (MT) system for a low-resource family of dialects, collectively known as Swiss German, which are widely spoken in Switzerland but seldom written. We collected a significant number of parallel written resources to start with, up to a total of about 60k words. Moreover, we identified several other promising data sources for Swiss German. Then, we designed and compared three strategies for normalizing Swiss German input in order to address the regional diversity. We found that character-based neural MT was the best solution for text normalization. In combination with phrase-based statistical MT, our solution reached 36% BLEU

score when translating from the Bernese dialect. This value, however, decreases as the testing data becomes more remote from the training one, geographically and topically. These resources and normalization techniques are a first step towards full MT of Swiss German dialects.

Improving domain-specific SMT for low-resourced languages using data from different domains

Fathima Farhath, Pranavan Theivendiram, Surangika Ranathunga, Sanath Jayasena and Gihan Dias

This paper evaluates the impact of different types of data sources in developing a domain-specific statistical machine translation (SMT) system for the domain of official government letters, between the low-resourced language pair Sinhala and Tamil. The baseline was built with a small in-domain parallel data set containing official government letters. The translation system was evaluated with two different test datasets. Test data from the same sources as training and tuning gave a higher score due to over-fitting, while the test data from a different source resulted in a considerably lower score. With the motive to improve translation, more data was collected from, (a) different government sources other than official letters (pseudo in-domain), and (b) online sources such as blogs, news and wiki dumps (out-domain). Use of pseudo in-domain data showed an improvement for both the test sets as the language is formal and context was similar to that of the in-domain though the writing style varies. Out-domain data, however, did not give a positive impact, either in filtered or unfiltered forms, as the writing style was different and the context was much more general than that of the official government documents.

Discovering Parallel Language Resources for Training MT Engines

Vassilis Papavassiliou, Prokopis Prokopidis and Stelios Piperidis

Web crawling is an efficient way for compiling the monolingual, parallel and/or domain-specific corpora needed for machine translation and other HLT applications. These corpora can be automatically processed to generate second order or synthesized derivative resources, including bilingual (general or domain-specific) lexica and terminology lists. In this submission, we discuss the architecture and use of the ILSP Focused Crawler (ILSP-FC), a system developed by researchers of the ILSP/Athena RIC for the acquisition of such resources, and currently being used through the European Language Resource Coordination effort. ELRC aims to identify and gather language and translation data relevant to public services and governmental institutions across 30 European countries participating in the Connecting Europe Facility (CEF).

A fine-grained error analysis of NMT, SMT and RBMT output for English-to-Dutch

Laura Van Brussel, Arda Tezcan and Lieve Macken

This paper presents a fine-grained error comparison of the English-to-Dutch translations of a commercial neural, phrase-based and rule-based machine translation (MT) system. For phrase-based and rule-based machine translation, we make use of the annotated SCATE corpus of MT errors, enriching it with the annotation of neural MT errors and updating the SCATE error taxonomy to fit the neural MT output as well. Neural, in general, outperforms phrase-based and rule-based systems especially for fluency, except for lexical issues. On the accuracy level, the improvements are less obvious. The target sentence does not always contain traces or clues of content being missing (omissions). This has repercussions for quality estimation or gisting operating only on the monolingual level. Mistranslations are part of another well represented error category, comprising a high number of word-sense disambiguation errors and a variety of other mistranslation errors, making it more complex to annotate or post-edit.

Collection and Analysis of Code-switch Egyptian Arabic-English Speech Corpus

Injy Hamed, Mohamed Elmahdy and Slim Abdennadher

Speech corpora are key components needed by both: linguists (in language analyses, research and teaching languages) and Natural Language Processing (NLP) researchers (in training and evaluating several NLP tasks such as speech recognition, text-to-speech and speech-to-text synthesis). Despite of the great demand, there is still a huge shortage in available corpora, especially in the case of dialectal languages, and code-switched speech. In this paper, we present our efforts in collecting and analyzing a speech corpus for conversational Egyptian Arabic. As in other multilingual societies, it is common among Egyptians to use a mix of Arabic and English in daily conversations. The act of switching languages, at sentence boundaries or within the same sentence, is referred to as code-switching. The aim of this work is a three-fold: (1) gather conversational Egyptian Arabic spontaneous speech, (2) obtain manual transcriptions and (3) analyze the speech from the code-switching perspective. A subset of the transcriptions were manually annotated for part-of-speech (POS) tags. The POS distribution of the embedded words was analyzed as well as the POS distribution for the trigger words (Arabic words preceding a code-switching point). The speech corpus can be obtained by contacting the authors.

Multimodal Lexical Translation

Chiraag Lala and Lucia Specia

Inspired by the tasks of Multimodal Machine Translation and Visual Sense Disambiguation we introduce a task called Multimodal Lexical Translation (MLT). The aim of this new task is to correctly translate an ambiguous word given its context - an image and a sentence in the source language. To facilitate the task, we introduce the MLT dataset, where each data point is a 4-tuple consisting of an ambiguous source word, its visual context (an image), its textual context (a source sentence), and its translation that conforms with the visual and textual contexts. The dataset has been created from the Multi30K corpus using word-alignment followed by human inspection for translations from English to German and English to French. We also introduce a simple heuristic to quantify the extent of the ambiguity of a word from the distribution of its translations and use it to select subsets of the MLT Dataset which are difficult to translate. These form a valuable multimodal and multilingual language resource with several potential uses including evaluation of lexical disambiguation within (Multimodal) Machine Translation systems.

Literality and cognitive effort: Japanese and Spanish

Isabel Lacruz, Michael Carl and Masaru Yamada

We introduce a notion of pause-word ratio computed using ranges of pause lengths rather than lower cutoffs for pause lengths. Standard pause-word ratios are indicators of cognitive effort during different translation modalities.The pause range version allows for the study of how different types of pauses relate to the extent of cognitive effort and where it occurs in the translation process. In this article we focus on short monitoring pauses and how they relate to the cognitive effort involved in translation and post-editing for language pairs that are different in terms of semantic and syntactic remoteness. We use data from the CRITT TPR database, comparing translation and post-editing from English to Japanese and from English to Spanish, and study the interaction of pause-word ratio for short pauses ranging between 300 and 500ms with syntactic remoteness, measured by the CrossS feature, semantic remoteness, measured by HTra, and syntactic and semantic remoteness, measured by Literality.

Evaluation of Machine Translation Performance Across Multiple Genres and Languages

Marlies Van der Wees, Arianna Bisazza and Christof Monz

In this paper, we present evaluation corpora covering four genres for four language pairs that we harvested from the web in an automated fashion. We use these multi-genre benchmarks to evaluate the impact of genre differences on machine translation (MT). We observe that BLEU score differences between genres can be large and that, for all genres and all language pairs, translation quality improves when using four genre-optimized systems rather than a single genre-agnostic system. Finally, we train and use genre classifiers to route test documents to the most appropriate genre systems. The results of these experiments show that our multi-genre benchmarks can serve to advance research on text genre adaptation for MT.

A Multilingual Dataset for Evaluating Parallel Sentence Extraction from Comparable Corpora

Pierre Zweigenbaum, Serge Sharoff and Reinhard Rapp

Comparable corpora can be seen as a reservoir for parallel sentences and phrases to overcome limitations in variety and quantity encountered in existing parallel corpora. This has motivated the design of methods to extract parallel sentences from comparable corporad. Despite this interest and work, no shared dataset has been made available for this task until the 2017 BUCC Shared Task. We present the challenges faced to build such a dataset and the solutions adopted to design and create the 2017 BUCC Shared Task dataset, emphasizing issues we had to cope with to include Chinese as one of the languages. The resulting corpus contains a total of about 3.5 million distinct sentences in English, French, German, Russian, and Chinese, mostly from Wikipedia. We illustrate the use of this dataset in the shared task and summarize the main results obtained by its participants. We finally outline remaining issues.

Manual vs Automatic Bitext Extraction

Aibek Makazhanov, Bagdat Myrzakhmetov and Zhenisbek Assylbekov

We compare manual and automatic approaches to the problem of extracting bitexts from the Web in the framework of a case study on building a Russian-Kazakh parallel corpus. Our findings suggest that targeted, site-specific crawling results in cleaner bitexts with a higher ratio of parallel sentences. We also find that general crawlers combined with boilerplate removal tools tend to retrieve shorter texts, as some content gets cleaned out with the markup. When it comes to sentence splitting and alignment we show that investing some effort in data pre- and post-processing as well as fiddling with off-the-shelf solutions pays a noticeable dividend. Overall we observe that, depending on the source, automatic bitext extraction methods may lack severely in coverage (retrieve fewer sentence pairs) and on average are less precise (retrieve fewer parallel sentence pairs). We conclude that if one aims at extracting high-quality bitexts for a small number of

language pairs, automatic methods best be avoided, or at least used with caution.

Session P50 - Morphology (2)

11th May 2018, 09:45

Chair person: **Amália Mendes** Poster Session

A Morphologically Annotated Corpus of Emirati Arabic

Salam Khalifa, Nizar Habash, Fadhl Eryani, Ossama Obeid, Dana Abdulrahim and Meera Al Kaabi

We present an ongoing effort on the first large-scale morphologically manually annotated corpus of Emirati Arabic. This corpus includes about 200,000 words selected from eight Gumar corpus novels from the Emirati Arabic variety. The selected texts are being annotated for tokenization, part-of-speech, lemmatization, English glosses and dialect identification. The orthography of the text is also adjusted for errors and inconsistencies. We discuss the guidelines for each part of the annotation components and the annotation interface we use. We report on the quality of the annotation through an inter annotator agreement measure.

CoNLL-UL: Universal Morphological Lattices for Universal Dependency Parsing

Amir More, Özlem Çetinoğlu, Çağrı Çöltekin, Nizar Habash, Benoît Sagot, Djamé Seddah, Dima Taji and Reut Tsarfaty

Following the development of the universal dependencies (UD) framework and the CoNLL 2017 Shared Task on end-to-end UD parsing, we address the need for a universal representation of morphological analysis which on the one hand can capture a range of different alternative morphological analyses of surface tokens, and on the other hand is compatible with the segmentation and morphological annotation guidelines prescribed for UD treebanks. We propose the CoNLL universal lattices (CoNLL-UL) format, a new annotation format for word lattices that represent morphological analyses, and provide resources that obey this format for a range of typologically different languages. The resources we provide are harmonized with the two-level representation and morphological annotation in their respective UD v2 treebanks, thus enabling research on universal models for morphological and syntactic parsing, in both pipeline and joint settings, and presenting new opportunities in the development of UD resources for low-resource languages.

Manually Annotated Corpus of Polish Texts Published between 1830 and 1918

Witold Kieraś and Marcin Woliński

The paper presents a manually annotated 625,000 tokens large historical corpus of – fiction, drama, popular science, essays and newspapers of the period. The corpus provides three layers: transliteration, transcription and morphosyntactic annotation. The annotation process as well as the corpus itself are described in detail in the paper.

Evaluating Inflectional Complexity Crosslinguistically: a Processing Perspective

Claudia Marzi, Marcello Ferro, Ouafae Nahli, Patrizia Belik, Stavros Bompolas and Vito Pirrelli

The paper provides a cognitively motivated method for evaluating the inflectional complexity of a language, based on a sample of "raw" inflected word forms processed and learned by a recurrent self-organising neural network with fixed parameter setting. Training items contain no information about either morphological content or structure. This makes the proposed method independent of both meta-linguistic issues (e.g. format and expressive power of descriptive rules, manual or automated segmentation of input forms, number of inflectional classes etc.) and language-specific typological aspects (e.g. word-based, stem-based or template-based morphology). Results are illustrated by contrasting Arabic, English, German, Greek, Italian and Spanish.

Parser combinators for Tigrinya and Oromo morphology

Patrick Littell, Tom McCoy, Na-Rae Han, Shruti Rijhwani, Zaid Sheikh, David R. Mortensen, Teruko Mitamura and Lori Levin

We present rule-based morphological parsers in the Tigrinya and Oromo languages, based on a parser-combinator rather than finite-state paradigm. This paradigm allows rapid development and ease of integration with other systems, although at the cost of non-optimal theoretical efficiency. These parsers produce multiple output representations simultaneously, including lemmatization, morphological segmentation, and an English word-for-word gloss, and we evaluate these representations as input for entity detection and linking and humanitarian need detection.

Massively Translingual Compound Analysis and Translation Discovery

Winston Wu and David Yarowsky

Word formation via compounding is a very widely observed but quite diverse phenomenon across the world's languages, but the

compositional semantics of a compound are often productively correlated between even distant languages. Using only freely available bilingual dictionaries and no annotated training data, we derive novel models for analyzing compound words and effectively generate novel foreign-language translations of English concepts using these models. In addition, we release a massively multilingual dataset of compound words along with their decompositions, covering over 21,000 instances in 329 languages, a previously unprecedented scale which should both productively support machine translation (especially in low resource languages) and also facilitate researchers in their further analysis and modeling of compounds and compound processes across the world's languages.

Building a Morphological Treebank for German from a Linguistic Database

Petra Steiner and Josef Ruppenhofer

German is a language with complex morphological processes. Its long and often ambiguous word forms present a bottleneck problem in natural language processing. As a step towards morphological analyses of high quality, this paper introduces a morphological treebank for German. It is derived from the linguistic database CELEX which is a standard resource for German morphology. We build on its refurbished, modernized and partially revised version. The derivation of the morphological trees is not trivial, especially for such cases of conversions which are morpho-semantically opaque and merely of diachronic interest. We develop solutions and present exemplary analyses. The resulting database comprises about 40,000 morphological trees of a German base vocabulary whose format and grade of detail can be chosen according to the requirements of the applications. The Perl scripts for the generation of the treebank are publicly available on github. In our discussion, we show some future directions for morphological treebanks. In particular, we aim at the combination with other reliable lexical resources such as GermaNet.

Session P51 - Multilinguality

11th May 2018, 09:45

Chair person: **Pavel Straňák** Poster Session

Baselines and Test Data for Cross-Lingual Inference

Željko Agić and Natalie Schluter

The recent years have seen a revival of interest in textual entailment, sparked by i) the emergence of powerful deep neural network learners for natural language processing and ii) the timely development of large-scale evaluation datasets such as SNLI. Recast as natural language inference, the problem now amounts to detecting the relation between pairs of statements: they either contradict or entail one another, or they are mutually neutral. Current research in natural language inference is effectively exclusive to English. In this paper, we propose to advance the research in SNLI-style natural language inference toward multilingual evaluation. To that end, we provide test data for four major languages: Arabic, French, Spanish, and Russian. We experiment with a set of baselines. Our systems are based on cross-lingual word embeddings and machine translation. While our best system scores an average accuracy of just over 75%, we focus largely on enabling further research in multilingual inference.

CATS: A Tool for Customized Alignment of Text Simplification Corpora

Sanja Štajner, Marc Franco-Salvador, Paolo Rosso and Simone Paolo Ponzetto

In text simplification (TS), parallel corpora consisting of original sentences and their manually simplified counterparts are very scarce and small in size, which impedes building supervised automated TS systems with sufficient coverage. Furthermore, the existing corpora usually do not distinguish sentence pairs which present full matches (both sentences contain the same information), and those that present only partial matches (the two sentences share the meaning only partially), thus not allowing for building customized automated TS systems which would separately model different simplification transformations. In this paper, we present our freely available, language-independent tool for sentence alignment from parallel/comparable TS resources (document-aligned resources), which additionally offers the possibility for filtering sentences depending on the level of their semantic overlap. We perform in-depth human evaluation of the tool's performance on English and Spanish corpora, and explore its capacities for classification of sentence pairs according to the simplification operation they model.

KIT-Multi: A Translation-Oriented Multilingual Embedding Corpus

Thanh-Le Ha, Jan Niehues, Matthias Sperber, Ngoc Quan Pham and Alexander Waibel

Cross-lingual word embeddings are the representations of words across languages in a shared continuous vector space. Cross-lingual word embeddings have been shown to be helpful in the development of cross-lingual natural language processing tools. In case of more than two languages involved, we call them multilingual word embeddings. In this work, we introduce a

multilingual word embedding corpus which is acquired by using neural machine translation. Unlike other cross-lingual embedding corpora, the embeddings can be learned from significantly smaller portions of data and for multiple languages at once. An intrinsic evaluation on monolingual tasks shows that our method is fairly competitive to the prevalent methods but on the cross-lingual document classification task, it obtains the best figures. Furthermore, the corpus is being analyzed regarding its usage and usefulness in other cross-lingual tasks. \\ \newline \Keywords{multilingual embeddings, cross-lingual embeddings, neural machine translation, multi-source translation} }

Multi-lingual Argumentative Corpora in English, Turkish, Greek, Albanian, Croatian, Serbian, Macedonian, Bulgarian, Romanian and Arabic

Alfred Sliwa, Yuan Man, Ruishen Liu, Niravkumar Borad, Seyedeh Ziyaei, Mina Ghobadi, Firas Sabbah and Ahmet Aker

Argumentative corpora are costly to create and are available in only few languages with English dominating the area. In this paper we release the first publicly available corpora in all Balkan languages and Arabic. The corpora are obtained by using parallel corpora where the source language is English and target language is either a Balkan language or Arabic. We use 8 different argument mining classifiers trained for English, apply them all on the source language and project the decision made by the classifiers to the target language. We assess the performance of the classifiers on a manually annotated news corpus. Our results show when at least 3 to 6 classifiers are used to judge a piece of text as argumentative an F1-score above 90% is obtained.

SemR-11: A Multi-Lingual Gold-Standard for Semantic Similarity and Relatedness for Eleven Languages

Siamak Barzegar, Brian Davis, Manel Zarrouk, Siegfried Handschuh and André Freitas

This work describes SemR-11, a multi-lingual dataset for evaluating semantic similarity and relatedness for 11 languages (German, French, Russian, Italian, Dutch, Chinese, Portuguese, Swedish, Spanish, Arabic and Persian). Semantic similarity and relatedness gold standards have been initially used to support the evaluation of semantic distance measures in the context of linguistic and knowledge resources and distributional semantic models. SemR-11 builds upon the English gold-standards of Miller & Charles (MC), Rubenstein & Goodenough (RG), WordSimilarity 353 (WS-353), and Simlex-999, providing a canonical translation for them. The final dataset consists of 15,917 word pairs and can be used to support the construction and evaluation of semantic similarity/relatedness and distributional semantic models. As a case study, the SemR-11 test collections was used to investigate how different distributional semantic models built from corpora in different languages and with different sizes perform in computing semantic relatedness similarity and relatedness tasks.

Session P52 - Part-of-Speech Tagging

11th May 2018, 09:45

Chair person: **Andreas Witt** Poster Session

Corpora with Part-of-Speech Annotations for Three Regional Languages of France: Alsatian, Occitan and Picard

Delphine Bernhard, Anne-Laure Ligozat, Fanny Martin, Myriam Bras, Pierre Magistry, Marianne Vergez-Couret, Lucie Steiblé, Pascale Erhart, Nabil Hathout, Dominique Huck, Christophe Rey, Philippe Reynés, Sophie Rosset, Jean Sibille and Thomas Lavergne

This article describes the creation of corpora with part-of-speech annotations for three regional languages of France: Alsatian, Occitan and Picard. These manual annotations were performed in the context of the RESTAURE project, whose goal is to develop resources and tools for these under-resourced French regional languages. The article presents the tagsets used in the annotation process as well as the resulting annotated corpora.

Part-of-Speech Tagging for Arabic Gulf Dialect Using Bi-LSTM

Randah Alharbi, Walid Magdy, Kareem Darwish, Ahmed Abdelali and Hamdy Mubarak

Part-of-speech (POS) tagging is one of the most important addressed areas in the natural language processing (NLP). There are effective POS taggers for many languages including Arabic. However, POS research for Arabic focused mainly on Modern Standard Arabic (MSA), while less attention was directed towards Dialect Arabic (DA). MSA is the formal variant which is mainly found in news and formal text books, while DA is the informal spoken Arabic that varies among different regions in the Arab world. DA is heavily used online due to the large spread of social media, which increased research directions towards building NLP tools for DA. Most research on DA focuses on Egyptian and Levantine, while much less attention is given to the Gulf dialect. In this paper, we present a more effective POS tagger for the Arabic Gulf dialect than currently available Arabic POS taggers. Our work includes preparing a POS tagging dataset, engineering multiple sets of features, and applying two machine learning

methods, namely Support Vector Machine (SVM) classifier and bi-directional Long Short Term Memory (Bi-LSTM) for sequence modeling. We have improved POS tagging for Gulf dialect from 75% accuracy using a state-of-the-art MSA POS tagger to over 91% accuracy using a Bi-LSTM labeler.

Web-based Annotation Tool for Inflectional Language Resources

Abdulrahman Alosaimy and Eric Atwell

We present Wasim, a web-based tool for semi-automatic morphosyntactic annotation of inflectional languages resources. The tool features high flexibility in segmenting tokens, editing, diacritizing, and labelling tokens and segments. Text annotation of highly inflectional languages (including Arabic) requires key functionality which we could not see in a survey of existing tools. Wasim integrates with morphological analysers to speed up the annotation process by selecting one from their proposed analyses. It integrates as well with external POS taggers for kick-start annotation and adaptive predicting based on annotations made so far. It aims to speed up the annotation by completely relying on a keyboard interface, with no mouse interaction required. Wasim has been tested on four case studies and these features proved to be useful. The source-code is released under the MIT license.

HiNTS: A Tagset for Middle Low German

Fabian Barteld, Sarah Ihden, Katharina Dreessen and Ingrid Schröder

In this paper, we describe the "Historisches Niederdeutsch Tagset" (HiNTS). This tagset has been developed for annotating parts-of-speech and morphology in Middle Low German texts, a group of historical (1200–1650) dialects of German. A non-standardized language such as Middle Low German has special conditions and requirements which have to be considered when designing a tagset for part of speech and morphology. We explain these requirements, i.e. the need to encode ambiguities while allowing the annotator to be as specific as possible, and our approach for dealing with them in the tagset. We then describe two special features of the tagset. In order to prove the benefit of these tags and corresponding annotation rules, we present example searches and the possible analyses arising from the results of such searches. Besides the usefulness of our tagset, we also considered its reliability in annotation using inter-annotator agreement experiments. The results of these experiments are presented and explained.

Leveraging Lexical Resources and Constraint Grammar for Rule-Based Part-of-Speech Tagging in Welsh

Steven Neale, Kevin Donnelly, Gareth Watkins and Dawn Knight

As the quantity of annotated language data and the quality of machine learning algorithms have increased over time, statistical part-of-speech (POS) taggers trained over large datasets have become as robust or better than their rule-based counterparts. However, for lesser-resourced languages such as Welsh there is simply not enough accurately annotated data to train a statistical POS tagger. Furthermore, many of the more popular rule-based taggers still require that their rules be inferred from annotated data, which while not as extensive as that required for training a statistical tagger must still be sizeable. In this paper we describe CyTag, a rule-based POS tagger for Welsh based on the VISL Constraint Grammar parser. Leveraging lexical information from Eurfa (an open-source dictionary for Welsh), we extract lists of possible POS tags for each word token in a running text and then apply various constraints - to prune the number of possible tags until the most appropriate tag for a given token can be selected. We explain how this approach is particularly useful in dealing with some of the specific intricacies of Welsh and present an evaluation of the performance of the tagger using a manually checked test corpus of 611 Welsh sentences.

Graph Based Semi-Supervised Learning Approach for Tamil POS tagging

Mokanarangan Thayaparan, Surangika Ranathunga and Uthayasanker Thayasivam

Parts of Speech (POS) tagging is an important pre-requisite for various Natural Language Processing tasks. POS tagging is rather challenging for morphologically rich languages such as Tamil. Being low-resourced, Tamil does not have a large POS annotated corpus to build good quality POS taggers using supervised machine learning techniques. In order to gain the maximum out of the existing Tamil POS tagged corpora, we have developed a graph-based semi-supervised learning approach to classify unlabelled data by exploiting a small sized POS labelled data set. In this approach, both labelled and unlabelled data are converted to vectors using word embeddings and a weighted graph is constructed using Mahalanobis distance. Then semi-supervised learning (SSL) algorithms are used to classify the unlabelled data. We were able to gain an accuracy of 0.8743 over an accuracy of 0.7333 produced by a CRF tagger for the same limited size corpus.

Session O33 - Lexicon

11th May 2018, 11:45

Chair person: **Simon Krek** Oral Session

The MADAR Arabic Dialect Corpus and Lexicon

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann and Kemal Oflazer

In this paper, we present two resources that were created as part of the Multi Arabic Dialect Applications and Resources (MADAR) project. The first is a large parallel corpus of 25 Arabic city dialects in the travel domain. The second is a lexicon of 1,045 concepts with an average of 45 words from 25 cities per concept. These resources are the first of their kind in terms of the breadth of their coverage and the fine location granularity. The focus on cities, as opposed to regions in studying Arabic dialects, opens new avenues to many areas of research from dialectology to dialect identification and machine translation.

Designing a Collaborative Process to Create Bilingual Dictionaries of Indonesian Ethnic Languages

Arbi Haza Nasution, Yohei Murakami and Toru Ishida

The constraint-based approach has been proven useful for inducing bilingual dictionary for closely-related low-resource languages. When we want to create multiple bilingual dictionaries linking several languages, we need to consider manual creation by a native speaker if there are no available machine-readable dictionaries are available as input. To overcome the difficulty in planning the creation of bilingual dictionaries, the consideration of various methods and costs, plan optimization is essential. Utilizing both constraint-based approach and plan optimizer, we design a collaborative process for creating 10 bilingual dictionaries from every combination of 5 languages, i.e., Indonesian, Malay, Minangkabau, Javanese, and Sundanese. We further design an online collaborative dictionary generation to bridge spatial gap between native speakers. We define a heuristic plan that only utilizes manual investment by the native speaker to evaluate our optimal plan with total cost as an evaluation metric. The optimal plan outperformed the heuristic plan with a 63.3% cost reduction.

Constructing a Lexicon of Relational Nouns

Edward Newell and Jackie Chi Kit Cheung

Relational nouns refer to an entity by virtue of how it relates to another entity. Their identification in text is a prerequisite for the correct semantic interpretation of a sentence, and could be used to improve information extraction. Although various systems for extracting relations expressed using nouns have been developed, there are no dedicated lexical resources for relational nouns. We contribute a lexicon of 6,224 labeled nouns which includes 1,446 relational nouns. We describe the bootstrapped annotation of relational nouns, and develop a classifier that achieves 70.4% F1 when tested on held out nouns that are among the most common 2,500 word types in Gigaword. We make the lexicon and classifier available to the scientific community.

Creating Large-Scale Multilingual Cognate Tables

Winston Wu and David Yarowsky

Low-resource languages often suffer from a lack of high-coverage lexical resources. In this paper, we propose a method to generate cognate tables by clustering words from existing lexical resources. We then employ character-based machine translation methods in solving the task of cognate chain completion by inducing missing word translations from lower-coverage dictionaries to fill gaps in the cognate chain, finding improvements over single language pair baselines when employing simple but novel multi-language system combination on the Romance and Turkic language families. For the Romance family, we show that system combination using the results of clustering outperforms weights derived from the historical-linguistic scholarship on language phylogenies. Our approach is applicable to any language family and has not been previously performed at such scale. The cognate tables are released to the research community.

Lexical Profiling of Environmental Corpora

Patrick Drouin, Marie-Claude L'Homme and Benoît Robichaud

This paper describes a method for distinguishing lexical layers in environmental corpora (i.e. the general lexicon, the transdisciplinary lexicon and two sets of lexical items related to the domain). More specifically we aim to identify the general environmental lexicon (GEL) and assess the extent to which we can set it apart from the others. The general intuition on which this research is based is that the GEL is both well-distributed in a specialized corpus (criterion 1) and specific to this type of corpora (criterion 2). The corpus used in the current experiment, made of 6 subcorpora that amount to 4.6 tokens, was compiled manually by terminologists for different projects designed to enrich a terminological resource. In order to meet criterion 1, the distribution of the GEL candidates is evaluated using a simple and well-known measure called. As for criterion 2, GEL candidates are extracted using a term extractor, which provides a measure of their specificity relative to a corpus. Our study focuses on

single-word lexical items including nouns, verbs and adjectives. The results were validated by a team of 4 annotators who are all familiar with the environmental lexicon and they show that using a high specificity threshold and a low idf threshold constitutes a good starting point to identify the GEL layer in our corpora.

Session O34 - Knowledge Discovery

11th May 2018, 11:45

Chair person: **German Rigau** Oral Session

Linking, Searching, and Visualizing Entities in Wikipedia

Marcus Klang and Pierre Nugues

In this paper, we describe a new system to extract, index, search, and visualize entities in Wikipedia. To carry out the entity extraction, we designed a high-performance, multilingual, entity linker and we used a document model to store the resulting linguistic annotations. The entity linker, HEDWIG, extracts the mentions from text using a string matching engine and links them to entities with a combination of statistical rules and PageRank. The document model, Docforia, consists of layers, where each layer is a sequence of ranges describing a specific annotation, here the entities. We evaluated HEDWIG with the TAC 2016 data and protocol and we reached the CEAFm scores of 70.0 on English, on 64.4 on Chinese, and 66.5 on Spanish. We applied the entity linker to the whole collection of English and Swedish articles of Wikipedia and we used Lucene to index the layers and a search module to interactively retrieve all the concordances of an entity in Wikipedia. The user can select and visualize the concordances in the articles or paragraphs. Contrary to classic text indexing, this system does not use strings to identify the entities but unique identifiers from Wikidata. A demonstration of the entity search and visualization will be available for English at this address http://vilde.cs.lth.se:9001/en-hedwig/ and for Swedish at: http://vilde.cs.lth.se:9001/sv-hedwig/.

Learning to Map Natural Language Statements into Knowledge Base Representations for Knowledge Base Construction

Chin-Ho Lin, Hen-Hsen Huang and Hsin-Hsi Chen

Directly adding the knowledge triples obtained from open information extraction systems into a knowledge base is often impractical due to a vocabulary gap between natural language (NL) expressions and knowledge base (KB) representation. This paper aims at learning to map relational phrases in triples from natural-language-like statement to knowledge base predicate format. We train a word representation model on a vector space and link each NL relational pattern to the semantically equivalent KB predicate. Our mapping result shows not only high quality, but also promising coverage on relational phrases compared to previous research.

Building a Knowledge Graph from Natural Language Definitions for Interpretable Text Entailment Recognition

Vivian Silva, André Freitas and Siegfried Handschuh

Natural language definitions of terms can serve as a rich source of knowledge, but structuring them into a comprehensible semantic model is essential to enable them to be used in semantic interpretation tasks. We propose a method and provide a set of tools for automatically building a graph world knowledge base from natural language definitions. Adopting a conceptual model composed of a set of semantic roles for dictionary definitions, we trained a classifier for automatically labeling definitions, preparing the data to be later converted to a graph representation. WordNetGraph, a knowledge graph built out of noun and verb WordNet definitions according to this methodology, was successfully used in an interpretable text entailment recognition approach which uses paths in this graph to provide clear justifications for entailment decisions.

Combining rule-based and embedding-based approaches to normalize textual entities with an ontology

Arnaud Ferré, Louise Deléger, Pierre Zweigenbaum and Claire Nédellec

In this paper, we propose a two-step method to normalize multi-word terms with concepts from a domain-specific ontology. Normalization is a critical step of information extraction. The method uses vector representations of terms computed with word embedding information and hierarchical information among ontology concepts. A training dataset and a first result dataset with high precision and low recall are generated by using the ToMap unsupervised normalization method. It is based on the similarities between the form of the term to normalize and the form of concept labels. Then, a projection of the space of terms towards the space of concepts is learned by globally minimizing the distances between vectors of terms and vectors of concepts. It applies multivariate linear regression using the previously generated training dataset. Finally, a distance calculation is carried out between the projections of term vectors and the concept vectors, providing a prediction of normalization by a concept for each term. This method was evaluated through the categorization task of bacterial habitats of BioNLP Shared Task 2016. Our results largely outperform all existing systems on this task, opening up very encouraging prospects.

T-REx: A Large Scale Alignment of Natural Language with Knowledge Base Triples

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest and Elena Simperl

Alignments between natural language and Knowledge Base (KB) triples are an essential prerequisite for training machine learning approaches employed in a variety of Natural Language Processing problems. These include Relation Extraction, KB Population, Question Answering and Natural Language Generation from KB triples. Available datasets that provide those alignments are plagued by significant shortcomings – they are of limited size, they exhibit a restricted predicate coverage, and/or they are of unreported quality. To alleviate these shortcomings, we present T-REx, a dataset of large-scale alignments between Wikipedia abstracts and Wikidata triples.T-REx consists of 11 million triples aligned with 3.09 million Wikipedia abstracts (6.2 million sentences). T-REx is two orders of magnitude larger than the largest available alignments dataset and covers 2.5 times more predicates. Additionally, we stress the quality of this language resource thanks to an extensive crowdsourcing evaluation. T-REx is publicly available at https://w3id.org/t-rex.

Session O35 - Multilingual Corpora & Machine Translation

11th May 2018, 11:45

Chair person: **Eva Hajičová** Oral Session

Multilingual Parallel Corpus for Global Communication Plan

Kenji Imamura and Eiichiro Sumita

In this paper, we introduce the Global Communication Plan (GCP) Corpus, a multilingual parallel corpus being developed as part of the GCP. The GCP Corpus is intended to be develop speech translation systems; thus, it primarily consists of pseudo-dialogues between foreign visitors and local Japanese people. The GCP Corpus is sentence-aligned and covers four domains and ten languages, including many Asian languages. In this paper, we summarize the GCP and the current status of the GCP Corpus. Then, we describe some of the corpus' basic characteristics from the perspective of multilingual machine translation and compare direct, pivot, and zero-shot translation techniques.

A Large Parallel Corpus of Full-Text Scientific Articles

Felipe Soares, Viviane Moreira and Karin Becker

The Scielo database is an important source of scientific information in Latin America, containing articles from several research domains. A striking characteristic of Scielo is that many of its full-text contents are presented in more than one language, thus being a potential source of parallel corpora. In this article, we present the development of a parallel corpus from Scielo in three languages: English, Portuguese, and Spanish. Sentences were automatically aligned using the Hunalign algorithm for all language pairs, and for a subset of trilingual articles also. We demonstrate the capabilities of our corpus by training a Statistical Machine Translation system (Moses) for each language pair, which outperformed related works on scientific articles. Sentence alignment was also manually evaluated, presenting an average of 98.8% correctly aligned sentences across all languages. Our parallel corpus is freely available in the TMX format, with complementary information regarding article metadata.

NegPar: A parallel corpus annotated for negation

Qianchu Liu, Federico Fancellu and Bonnie Webber

Although the existence of English corpora annotated for negation has allowed for extensive work on monolingual negation detection, little is understood on how negation-related phenomena translate across languages. The current study fills this gap by presenting NegPar, the first English-Chinese parallel corpus annotated for negation in the narrative domain (a collection of stories from Conan Doyle's Sherlock Holmes). While we followed the annotation guidelines in the ConanDoyleNeg corpus (Morante and Daelemans, 2012), we reannotated certain scope-related phenomena to ensure more consistent and interpretable semantic representation. To both ease the annotation process and analyze how similar negation is signaled in the two languages, we experimented with first projecting the annotations from English and then manually correcting the projection output in Chinese. Results show that projecting negation via word-alignment offers limited help to the annotation process, as negation can be rendered in different ways across languages.

The IIT Bombay English-Hindi Parallel Corpus

Anoop Kunchukuttan, Pratik Mehta and Pushpak Bhattacharyya

We present the IIT Bombay English-Hindi Parallel Corpus. The corpus is a compilation of parallel corpora previously available in the public domain as well as new parallel corpora we collected. The corpus contains 1.49 million parallel segments, of which 694k segments were not previously available in the public domain. The corpus has been pre-processed for machine translation, and we report baseline phrase-based SMT and NMT translation results on this corpus. This corpus has been used in two editions of shared tasks at the Workshop on Asian Language Translation (2016 and 2017). The corpus is freely available for non-commercial

research. To the best of our knowledge, this is the largest publicly available English-Hindi parallel corpus.

Extracting an English-Persian Parallel Corpus from Comparable Corpora

Akbar Karimi, Ebrahim Ansari and Bahram Sadeghi Bigham

Parallel data are an important part of a reliable Statistical Machine Translation (SMT) system. The more of these data are available, the better the quality of the SMT system. However, for some language pairs such as Persian-English, parallel sources of this kind are scarce. In this paper, a bidirectional method is proposed to extract parallel sentences from English and Persian document aligned Wikipedia. Two machine translation systems are employed to translate from Persian to English and the reverse after which an IR system is used to measure the similarity of the translated sentences. Adding the extracted sentences to the training data of the existing SMT systems is shown to improve the quality of the translation. Furthermore, the proposed method slightly outperforms the one-directional approach. The extracted corpus consists of about 200,000 sentences which have been sorted by their degree of similarity calculated by the IR system and is freely available for public access on the Web.

Session O36 - Corpus Creation, Use & Evaluation (2)

11th May 2018, 11:45

Chair person: **Satoshi Nakamura** Oral Session

Learning Word Vectors for 157 Languages

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin and Tomas Mikolov

Distributed word representations, or word vectors, have recently been applied to many tasks in natural language processing, leading to state-of-the-art performance. A key ingredient to the successful application of these representations is to train them on very large corpora, and use these pre-trained models in downstream tasks. In this paper, we describe how we trained such high quality word representations for 157 languages. We used two sources of data to train these models: the free online encyclopedia Wikipedia and data from the common crawl project. We also introduce three new word analogy datasets to evaluate these word vectors, for French, Hindi and Polish. Finally, we evaluate our pre-trained word vectors on 10 languages for which evaluation datasets exists, showing very strong performance compared to previous models.

SumeCzech: Large Czech News-Based Summarization Dataset

Milan Straka, Nikita Mediankin, Tom Kocmi, Zdeněk Žabokrtský, Vojtěch Hudeček and Jan Hajic

Document summarization is a well-studied NLP task. With the emergence of artificial neural network models, the summarization performance is increasing, as are the requirements on training data. However, only a few datasets are available for Czech, none of them particularly large. Additionally, summarization has been evaluated predominantly on English, with the commonly used ROUGE metric being English-specific. In this paper, we try to address both issues. We present SumeCzech, a Czech news-based summarization dataset. It contains more than a million documents, each consisting of a headline, a several sentences long abstract and a full text. The dataset can be downloaded using the provided scripts available at http://hdl.handle.net/11234/1-2615. We evaluate several summarization baselines on the dataset, including a strong abstractive approach based on Transformer neural network architecture. The evaluation is performed using a language-agnostic variant of ROUGE.

A Diachronic Corpus for Literary Style Analysis

Carmen Klaussner and Carl Vogel

This research presents a resource for diachronic style analysis in particular the analysis of literary authors over time. Temporal style analysis has received comparatively little attention over the past years in spite of the possibility of an author's style frequently changing over time, a property that is not only interesting in its own right, but which also has implications for synchronic analyses of style. The corpus contains 22 American literary authors from the mid 19th to early 20th century that wrote largely in parallel. After describing the resource, we show how the corpus can be used to detect changing features in literary style.

Text Simplification from Professionally Produced Corpora

Carolina Scarton, Gustavo Paetzold and Lucia Specia

The lack of large and reliable datasets has been hindering progress in Text Simplification (TS). We investigate the application of the recently created Newsela corpus, the largest collection of professionally written simplifications available, in TS tasks. Using new alignment algorithms, we extract 550,644 complex-simple sentence pairs from the corpus. This data is explored in different ways: (i) we show that traditional readability metrics capture surprisingly well the different complexity levels in this corpus, (ii) we build machine learning models to classify sentences into complex vs. simple and to predict complexity levels

that outperform their respective baselines, (iii) we introduce a lexical simplifier that uses the corpus to generate candidate simplifications and outperforms the state of the art approaches, and (iv) we show that the corpus can be used to learn sentence simplification patterns in more effective ways than corpora used in previous work.

Intertextual Correspondence for Integrating Corpora

Jacky Visser, Rory Duthie, John Lawrence and Chris Reed

We present intertextual correspondence (ITC) as an integrative technique for combining annotated text corpora. The topical correspondence between different texts can be exploited to establish new annotation connections between existing corpora. Although the general idea should not be restricted to one particular theoretical framework, we explain how the annotation of intertextual correspondence works for two corpora annotated with argumentative notions on the basis of Inference Anchoring Theory. The annotated corpora we take as examples are topically and temporally related: the first corpus comprises television debates leading up to the 2016 presidential elections in the United States, the second corpus consists of commentary on and discussion of those debates on the social media platform Reddit. The integrative combination enriches the existing corpora in terms of the argumentative density, conceived of as the number of inference, conflict and rephrase relations relative to the word count of the (sub-)corpus. ITC also affects the global properties of the corpus, such as the most divisive issue. Moreover, the ability to extend existing corpora whilst maintaining the level of internal cohesion is beneficial to the use of the integrated corpus as resource for text and argument mining based on machine learning.

Session P53 - Conversational Systems/Dialogue/Chatbots/Human-Robot Interaction (3)

11th May 2018, 11:45

Chair person: **Kalika Bali** Poster Session

What Causes the Differences in Communication Styles? A Multicultural Study on Directness and Elaborateness

Juliana Miehle, Wolfgang Minker and Stefan Ultes

With the aim of designing a Spoken Dialogue System which adapts to the user's communication idiosyncrasies, we present a multicultural study to investigate the causes of differences in the communication styles elaborateness and directness in Human-Computer Interaction. By adapting the system's behaviour to the user, the conversation agent may appear more familiar and trustworthy. 339 persons from Germany, Russia, Poland, Spain and the United Kingdom participated in this web-based study. The participants had to imagine that they are talking to a digital agent. For every dialogue turn, they had to read four different variants of the system output and indicate their preference. With the results of this study, we could demonstrate the influence of the user's culture and gender, the frequency of use of speech based assistants as well as the system's role on the user's preference concerning the system's communication style in terms of its elaborateness and its directness.

FARMI: A FrAmework for Recording Multi-Modal Interactions

Patrik Jonell, Mattias Bystedt, Per Fallgren, Dimosthenis Kontogiorgos, José Lopes, Zofia Malisz, Samuel Mascarenhas, Catharine Oertel, Eran Raveh and Todd Shore

In this paper we present (1) a processing architecture used to collect multi-modal sensor data, both for corpora collection and real-time processing, (2) an open-source implementation thereof and (3) a use-case where we deploy the architecture in a multi-party deception game, featuring six human players and one robot. The architecture is agnostic to the choice of hardware (e.g. microphones, cameras, etc.) and programming languages, although our implementation is mostly written in Python. In our use-case, different methods of capturing verbal and non-verbal cues from the participants were used. These were processed in real-time and used to inform the robot about the participants' deceptive behaviour. The framework is of particular interest for researchers who are interested in the collection of multi-party, richly recorded corpora and the design of conversational systems. Moreover for researchers who are interested in human-robot interaction the available modules offer the possibility to easily create both autonomous and wizard-of-Oz interactions.

Creating Large-Scale Argumentation Structures for Dialogue Systems

Kazuki Sakai, Akari Inago, Ryuichiro Higashinaka, Yuichiro Yoshikawa, Hiroshi Ishiguro and Junji Tomita

We are planning to develop argumentative dialogue systems that can discuss various topics with people by using large-scale argumentation structures. In this paper, we describe the creation process of these argumentation structures. We created ten structures each having more than 2000 nodes of five topics in English and five topics in Japanese. We analyzed the created structures for their characteristics and investigated the differences between the two languages. We conducted an

evaluation experiment to ascertain that the structures can be applied to dialogue systems. We conducted another experiment to use the created argumentation structures as training data for augmenting the current argumentation structures.

Exploring Conversational Language Generation for Rich Content about Hotels

Marilyn Walker, Albry Smither, Shereen Oraby, Vrindavan Harrison and Hadar Shemtov

Dialogue systems for hotel and tourist information have typically simplified the richness of the domain, focusing system utterances on only a few selected attributes such as price, location and type of rooms. However, much more content is typically available for hotels, often as many as 50 distinct instantiated attributes for an individual entity. New methods are needed to use this content to generate natural dialogues for hotel information, and in general for any domain with such rich complex content. We describe three experiments aimed at collecting data that can inform an NLG for hotels dialogues, and show, not surprisingly, that the the sentences in the original written hotel descriptions provided on webpages for each hotel are stylistically not a very good match for conversational interaction. We quantify the stylistic features that characterize the differences between the original textual data and the collected dialogic data. We plan to use these in stylistic models for generation, and for scoring retrieved utterances for use in hotel dialogues.

Identification of Personal Information Shared in Chat-Oriented Dialogue

Sarah Fillwock and David Traum

We present an analysis of how personal information is shared in chat-oriented dialogue. We develop an annotation scheme, including entity-types, attributes, and values, that can be used to annotate the presence and type of personal information in these dialogues. A collection of attribute types is identified from the annotation of three chat-oriented dialogue corpora and a taxonomy of personal information pertinent to chat-oriented dialogue is presented. We examine similarities and differences in the frequency of specific attributes in the three corpora and observe that there is much overlap between the attribute types which are shared between dialogue participants in these different settings. The work presented here suggests that there is a common set of attribute types that frequently occur within chat-oriented dialogue in general. This resource can be used in the development of chat-oriented dialogue systems by providing common topics that a dialogue system should be able to talk about.

A Vietnamese Dialog Act Corpus Based on ISO 24617-2 standard

Thi Lan Ngo, Pham Khac Linh and Takeda Hideaki

The voice-based human-machine interaction systems such as personal virtual assistants, chat-bots, and automatic contact centres are becoming increasingly popular. In this trend, conversation mining research also is getting the attention of many researchers. Standardized data play an important role in conversation mining. In this paper, we present a new Vietnamese corpus annotated for dialog acts using the ISO 24617-2 standard (2012), for emotions using Ekman's six primitives (1972), and for sentiment using the tags "positive", "negative" and "neutral". Emotion and sentiment are tagged at functional segment level. We show how the corpus is constructed and provide a brief statistical description of the data. This is the first Vietnamese dialog act corpus.

Annotating Reflections for Health Behavior Change Therapy

Nishitha Guntakandla and Nielsen Rodney

We present our work on annotating reflections, an essential counselor behavioral code in motivational interviewing for psychotherapy on conversations that are a combination of casual and therapeutic dialogue. We annotated all the therapists' utterances from ten transcripts spanning more than five hours of conversation with Complex Reflection, Simple Reflection, or No Reflection. We also provide insights into corpus quality and code distributions. The corpus that we constructed and annotated in this effort is a vital resource for automated health behavior change therapy via a dialogue system. As the on-going work, additional conversations are being annotated at least by one annotator.

Session P54 - Discourse Annotation, Representation and Processing (2)

11th May 2018, 11:45

Chair person: **Bruno Cartoni** Poster Session

Annotating Attribution Relations in Arabic

Amal Alsaif, Tasniem Alyahya, Madawi Alotaibi, Huda Almuzaini and Abeer Algahtani

We present a first empirical effort in annotating attribution in Modern Standard Arabic (MSA). Identifying attributed arguments to the source is applied successfully in diverse systems such as authorship identification, information retrieval, and opinion mining. Current studies focus on using lexical terms in long texts to verify, for example, the author identity. While attribution identification in short texts is still unexplored completely due to the lack of resources such as annotated corpora and tools especially in Arabic on one hand, and the limited coverage of

different attribution usages in Arabic literature, on other hand. The paper presents our guidelines for annotating attribution elements: cue, source, and the content with required syntactical and semantic features in Arabic news (Arabic TreeBank - ATB) insight of earlier studies for other languages with all required adaptation. We also develop a new annotation tool for attribution in Arabic to ensure that all instances of attribution are reliably annotated. The results of a pilot annotation are discussed in addition to the inter-annotators agreement studies towards creating the first gold standard attribution corpus for Arabic.

The ADELE Corpus of Dyadic Social Text Conversations:Dialog Act Annotation with ISO 24617-2

Emer Gilmartin, Christian Saam, Brendan Spillane, Maria O'Reilly, Ketong Su, Arturo Calvo, Loredana Cerrato, Killian Levacher, Nick Campbell and Vincent Wade

Social or interactional dialog is less well described than task-based or instrumental dialog, although there is increasing interest in the genre, particularly in light of new spoken and text dialog applications which aim to relate to the user as well as perform tasks. Dialog act annotation aids understanding of interaction structure; essential to the design of sucessful artificial dialog. Much social text interaction may be closer to social talk than to traditional written language. In this paper we briefly describe social or casual talk, and review how current dialog annotation schemes and particularly the ISO standard 24617-2 (Semantic annotation framework, Part 2: Dialogue Acts) treat non-task elements of dialog. To aid in training a casual talk system, we collected a corpus of 193 dyadic text dialogs, based on a novel 'getting to know you' social dialog elicitation paradigm. We describe the annotation of the dialogs, and propose additional acts to better cover greeting and leavetaking. We report on preliminary analyses of the corpus, and provide an overview of the corpus content and its relationship to spoken language. The corpus, coding manual, and annotations are being packaged and will be made available to interested researchers.

An Assessment of Explicit Inter- and Intra-sentential Discourse Connectives in Turkish Discourse Bank

Deniz Zeyrek and Murathan Kurfalı

The paper offers a quantitative and qualitative analysis of explicit inter- and intra-sentential discourse connectives in Turkish Discourse Bank, or TDB version 1.1, a multi-genre resource of written Turkish manually annotated at the discourse level following the goals and principles of Penn Discourse TreeBank. TDB 1.1 is a 40K-word corpus involving all major discourse relation types (explicit discourse relations at intra- and inter-sentential positions, implicit discourse relations, alternative lexicalizations and entity relations) along with their senses and the text spans they relate. The paper focuses on the addition of a new set of explicit intra-sentential connectives to TDB 1.1, namely converbs (a subset of subordinators), which are suffixal connectives mostly corresponding to subordinating conjunctions in European languages. An evaluation of the converb sense annotations is provided. Then, with corpus statistics, explicit intra- and inter-sentential connectives are compared in terms of their frequency of occurrence and with respect to the senses they convey. The results suggest that the subordinators tend to select certain senses not selected by explicit inter-sentential discourse connectives in the data. Overall, our findings offer a promising direction for future NLP tasks in Turkish.

Compilation of Corpora for the Study of the Information Structure–Prosody Interface

Alicia Burga, Monica Dominguez, Mireia Farrús and Leo Wanner

Theoretical studies on the Information Structure–prosody interface argue that the content packaged in terms of theme and rheme correlates with the intonation of the corresponding sentence. However, there are few empirical studies that support this argument and even fewer resources that promote reproducibility and scalability of experiments. In this paper, we introduce a methodology for the compilation of annotated corpora to study the correspondence between Information Structure and prosody. The application of this methodology is exemplified on a corpus of read speech in English annotated with hierarchical thematicity and automatically extracted prosodic parameters.

Preliminary Analysis of Embodied Interactions between Science Communicators and Visitors Based on a Multimodal Corpus of Japanese Conversations in a Science Museum

Rui Sakaida, Ryosaku Makino and Mayumi Bono

This paper introduces preliminary analyses of embodied interactions, drawing on a multimodal corpus of Japanese conversations, which we video-recorded during scientific communications at a museum in Tokyo, the Miraikan. A comparison of similar cases extracted from our multimodal corpus shows both similarities and differences, not only in language use but also in bodily conduct in certain interactional sequences. We focus on a number of sequences, such as those where science communicators invite visitors to walk to the next exhibit, and our detailed analyses show that the practices of science communicators are context-free and context-sensitive

interactional procedures, adapted and adjusted to the different situations communicators may encounter. After presenting our analyses, based on a corpus from a naturally occurring but partly controlled setting, we suggest that we can investigate both the generality and the situatedness of interactional practices. In the future, using such multimodal corpora, we will be able to both qualitatively and quantitatively analyze language use and non-verbal behaviors in situated activities.

Improving Crowdsourcing-Based Annotation of Japanese Discourse Relations

Yudai Kishimoto, Shinnosuke Sawada, Yugo Murawaki, Daisuke Kawahara and Sadao Kurohashi

Although discourse parsing is an important and fundamental task in natural language processing, few languages have corpora annotated with discourse relations and if any, they are small in size. Creating a new corpus of discourse relations by hand is costly and time-consuming. To cope with this problem, Kawahara et al. (2014) constructed a Japanese corpus with discourse annotations through crowdsourcing. However, they did not evaluate the quality of the annotation. In this paper, we evaluate the quality of the annotation using expert annotations. We find out that crowdsourcing-based annotation still leaves much room for improvement. Based on the error analysis, we propose improvement techniques based on language tests. We re-annotated the corpus with discourse annotations using the improvement techniques, and achieved approximately 3% improvement in F-measure. We will make re-annotated data publicly available.

Persian Discourse Treebank and coreference corpus

Azadeh Mirzaei and Pegah Safari

This research addresses the investigation of intra-document relations based on two major approaches: discourse analysis and coreference resolution which results in building the first Persian discourse Treebank and a comprehensive Persian coreference corpus. In discourse analysis, we have explored sentence-level relations defined between clauses in complex sentences. So we specified 34682 discourse relations, the sense of the relations, their arguments and their attributes mainly consisted of the source of the message and its type. Our discourse analysis is based on a corpus consisted of 30000 individual sentences with morphologic, syntactic and semantic labels and nearly half a million tokens. Also 18336 of these sentences are double-annotated. For coreference annotation, since a document-based corpus was needed, we prepared a new corpus consisted of 547 documents and 212646 tokens which is still under development. We enriched it with morphological and syntactical labels and added coreference information at the top. Currently, we have annotated 6511 coreference chains and 21303 mentions with a comprehensive annotation scheme to compensate some specification of Persian such as being pro-drop or lacking gender agreement information.

Automatic Labeling of Problem-Solving Dialogues for Computational Microgenetic Learning Analytics

Yuanliang Meng, Anna Rumshisky and Florence Sullivan

This paper presents a recurrent neural network model to automate the analysis of students' computational thinking in problem-solving dialogue. We have collected and annotated dialogue transcripts from middle school students solving a robotics challenge, and each dialogue turn is assigned a code. We use sentence embeddings and speaker identities as features, and experiment with linear chain CRFs and RNNs with a CRF layer (LSTM-CRF). Both the linear chain CRF model and the LSTM-CRF model outperform the naive baselines by a large margin, and LSTM-CRF has an edge between the two. To our knowledge, this is the first study on dialogue segment annotation using neural network models. This study is also a stepping-stone to automating the microgenetic analysis of cognitive interactions between students.

Increasing Argument Annotation Reproducibility by Using Inter-annotator Agreement to Improve Guidelines

Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany and Serena Villata

In this abstract we present a methodology to improve Argument annotation guidelines by exploiting inter-annotator agreement measures. After a first stage of the annotation effort, we have detected problematic issues via an analysis of inter-annotator agreement. We have detected ill-defined concepts, which we have addressed by redefining high-level annotation goals. For other concepts, that are well-delimited but complex, the annotation protocol has been extended and detailed. Moreover, as can be expected, we show that distinctions where human annotators have less agreement are also those where automatic analyzers perform worse. Thus, the reproducibility of results of Argument Mining systems can be addressed by improving inter-annotator agreement in the training material. Following this methodology, we are enhancing a corpus annotated with argumentation, available at https://github.com/PLN-FaMAF/ArgumentMiningECHR together with guidelines and analyses of agreement. These analyses can be used to filter performance figures of automated

systems, with lower penalties for cases where human annotators agree less.

Session P55 - Language Acquisition & CALL (2)

11th May 2018, 11:45

Chair person: **Serge Sharoff** Poster Session

Semi-Supervised Clustering for Short Answer Scoring

Andrea Horbach and Manfred Pinkal

This paper investigates the use of semi-supervised clustering for Short Answer Scoring (SAS). In SAS, clustering techniques are an attractive alternative to classification because they provide structured groups of answers in addition to a score. Previous approaches use unsupervised clustering and have teachers label some items after clustering. We propose to re-allocate some of the human annotation effort to before and during the clustering process for (i) feature selection, (ii) for creating pairwise constraints and (iii) for metric learning. Our methods improve clustering performance substantially from 0.504 kappa for unsupervised clustering to 0.566.

Analyzing Vocabulary Commonality Index Using Large-scaled Database of Child Language Development

Yan Cao, Yasuhiro Minami, Yuko Okumura and Tessei Kobayashi

The present study proposed a vocabulary commonality index for child language development to investigate to what extent each child acquires common words during the early stages of lexical development. We used large-scaled, vocabulary-checklist data from Japanese-speaking children (N=1,451) aged 8-48 months to estimate their age of acquisition (AoA) of 2688 words by logistic regression. Then we calculated the vocabulary commonality index for each child with two datasets. The results showed that as their vocabulary size increases, children who have the same vocabulary size tend to produce common words with the same ratio.

The ICoN Corpus of Academic Written Italian (L1 and L2)

Mirko Tavosanis and Federica Cominetti

This paper describes the ICoN corpus, a corpus of academic written Italian, some of the directions of research it could open, and some of the first outcomes of research conducted on it. The ICoN corpus includes 2,115,000 tokens written by students having Italian as L2 students (level B2 or higher) and 1,769,000 tokens written by students having Italian as L1; this makes it the largest corpus of its kind. The texts included in the corpus come from the online examinations taken by 787 different students for the ICoN Degree Program in Italian Language and Culture for foreign students and Italian citizens residing abroad. The texts were produced by students having 41 different L1s, and 18 different L1s are represented in the corpus by more than 20,000 tokens. The corpus is encoded in XML files; it can be freely queried online and it is available upon request for research purposes. The paper includes the discussion of preliminary research in the field of collocations, showing that, in the texts included in the corpus, while learners and natives do use multiword expressions in a similar way, learners can overuse relatively infrequent forms of multiword adverbials, or use some adverbials in a non-standard way.

Revita: a Language-learning Platform at the Intersection of ITS and CALL

Anisia Katinskaia, Javad Nouri and Roman Yangarber

This paper presents Revita, a Web-based platform for language learning—beyond the beginner level. We anchor the presentation in a survey, where we review the literature about recent advances in the fields of computer-aided language learning (CALL) and intelligent tutoring systems (ITS). We outline the established desiderata of CALL and ITS and discuss how Revita addresses (the majority of) the theoretical requirements of CALL and ITS. Finally, we claim that, to the best of our knowledge, Revita is currently the only platform for learning/tutoring beyond the beginner level, that is functional, freely-available and supports multiple languages.

The Distribution and Prosodic Realization of Verb Forms in German Infant-Directed Speech

Bettina Braun and Katharina Zahner

Infant-directed speech is often seen as a predictor for infants' speech processing abilities, for instance speech segmentation or word learning. In this paper, we examine the syntactic distribution (position), accentuation and prosodic phrasing of German verb forms and discuss that many verb forms are prime candidates for early segmentation: they frequently appear at the start or end of prosodic phrases; if they are not phrase-initial, they are often preceded by closed-class word forms and they are frequently accented (imperative verb forms: 72% of the cases, infinitive verb forms: 82% of the cases). It thus appears that German infants ought to be able to extract verbs as early as nouns, given appropriate stimulus materials.

Cross-linguistically Small World Networks are Ubiquitous in Child-directed Speech

Steven Moran, Danica Pajović and Sabine Stoll

In this paper we use network theory to model graphs of child-directed speech from caregivers of children from nine typologically and morphologically diverse languages. With the resulting lexical adjacency graphs, we calculate the network statistics {N, E, <k>, L, C} and compare them against the standard baseline of the same parameters from randomly generated networks of the same size. We show that typologically and morphologically diverse languages all share small world properties in their child-directed speech. Our results add to the repertoire of universal distributional patterns found in the input to children cross-linguistically. We discuss briefly some implications for language acquisition research.

L1-L2 Parallel Treebank of Learner Chinese: Overused and Underused Syntactic Structures

Keying Li and John Lee

We present a preliminary analysis on a corpus of texts written by learners of Chinese as a foreign language (CFL), annotated in the form of an L1-L2 parallel dependency treebank. The treebank consists of parse trees of sentences written by CFL learners ("L2 sentences"), parse trees of their target hypotheses ("L1 sentences"), and word alignment between the L1 sentences and L2 sentences. Currently, the treebank consists of 600 L2 sentences and 697 L1 sentences. We report the most overused and underused syntactic relations by the CFL learners, and discuss the underlying learner errors.

The Use of Text Alignment in Semi-Automatic Error Analysis: Use Case in the Development of the Corpus of the Latvian Language Learners

Roberts Darģis, Ilze Auziņa and Kristīne Levāne-Petrova

This article presents a different method for creation of error annotated corpora. The approach suggested in this paper consists of multiple parts - text correction, automated morphological analysis, automated text alignment and error annotation. Error annotation can easily be semi-automated with a rule-based system, similar to the one used in this paper. The text correction can also be semi-automated using a rule-based system or even machine learning. The use of the text correction, word, and letter alignment enables more in-depth analysis of errors types, providing opportunities for quantitative research. The proposed method has been approbated in the development of the corpus of the Latvian language learners. Spelling, punctuation, grammatical, syntactic and lexical errors are annotated in the corpus. Text that is not understandable is marked as unclear for additional analysis. The method can easily be adapted for the development of error corpora in any other languages with relatively free word order. The highest gain from this method will be for highly inflected languages with rich morphology.

Error annotation in a Learner Corpus of Portuguese

Iria Del Río Gayo and Amália Mendes

We present the error tagging system of the COPLE2 corpus and the first results of its implementation.. The system takes advantage of the corpus architecture and the possibilities of the TEITOK environment to reduce manual effort and produce a final standoff, multi-level annotation with position-based tags that account for the main error types observed in the corpus. The first step of the tagging process involves the manual annotation of errors at the token level. We have already annotated 47% of the corpus using this approach. In a further step, the token-based annotations will be automatically transformed (fully or partially) in position-based error tags. COPLE2 is the first Portuguese learner corpus with error annotation. We expect that this work will support new research in different fields connected with Portuguese as second/foreign language, like Second Language Acquisition/Teaching or Computer Assisted Learning.

An SLA Corpus Annotated with Pedagogically Relevant Grammatical Structures

Leonardo Zilio, Rodrigo Wilkens and Cédrick Fairon

The evaluation of a language learner's proficiency in second language is a task that normally involves comparing the learner's production with a learning framework of the target language. A broad framework is the Common European Framework for Languages (CEFR), which addresses language learning in general and is broadly used in the European Union, while serving as reference in countries outside the EU as well. In this study, we automatically annotated a corpus of texts produced by language learners with pedagogically relevant grammatical structures and we observed how these structures are being employed by learners from different proficiency levels. We analyzed the use of structures both in terms of evolution along the levels and in terms of level in which the structures are used the most. The annotated resource, SGATe, presents a rich source of information for teachers that wish to compare the production of their students with those of already certified language learners.

Session P56 - Less-Resourced/Endangered Languages (2)

11th May 2018, 11:45

Chair person: **Sonja Bosch** Poster Session

Portable Spelling Corrector for a Less-Resourced Language: Amharic

Andargachew Mekonnen Gezmu, Andreas Nürnberger and Binyam Ephrem Seyoum

This paper describes an automatic spelling corrector for Amharic, the working language of the Federal Government of Ethiopia. We used a corpus-driven approach with the noisy channel for spelling correction. It infers linguistic knowledge from a text corpus. The approach can be ported to other written languages with little effort as long as they are typed using a QWERTY keyboard with direct mappings between keystrokes and characters. Since Amharic letters are syllabic, we used a modified version of the System for Ethiopic Representation in ASCII for transliteration in the like manner as most Amharic keyboard input methods do. The proposed approach is evaluated with Amharic and English test data and has scored better performance result than the baseline systems: GNU Aspell and Hunspell. We get better result due to the smoothed language model, the generalized error model and the ability to take into account the context of misspellings. Besides, instead of using a handcrafted lexicon for spelling error detection, we used a term list derived from frequently occurring terms in a text corpus. Such a term list, in addition to ease of compilation, has also an advantage in handling rare terms, proper nouns, and neologisms.

A Speaking Atlas of the Regional Languages of France

Philippe Boula de Mareüil, Albert Rilliard and Frédéric Vernier

The aim is to show and promote the linguistic diversity of France, through field recordings, a computer program (which allows us to visualise dialectal areas) and an orthographic transcription (which represents an object of research in itself). A website is presented (https://atlas.limsi.fr), displaying an interactive map of France from which Aesop's fable "The North Wind and the Sun" can be listened to and read in French and in 140 varieties of regional languages. There is thus both a scientific dimension and a heritage dimension in this work, insofar as a number of regional or minority languages are in a critical situation.

Towards Language Technology for Mi'kmaq

Anant Maheshwari, Leo Bouscarrat and Paul Cook

Mi'kmaq is a polysynthetic Indigenous language spoken primarily in Eastern Canada, on which no prior computational work has focused. In this paper we first construct and analyze a web corpus of Mi'kmaq. We then evaluate several approaches to language modelling for Mi'kmaq, including character-level models that are particularly well-suited to morphologically-rich languages. Preservation of Indigenous languages is particularly important in the current Canadian context; we argue that natural language processing could aid such efforts.

Pronunciation Dictionaries for the Alsatian Dialects to Analyze Spelling and Phonetic Variation

Lucie Steiblé and Delphine Bernhard

This article presents new pronunciation dictionaries for the under-resourced Alsatian dialects, spoken in north-eastern France. These dictionaries are compared with existing phonetic transcriptions of Alsatian, German and French in order to analyze the relationship between speech and writing. The Alsatian dialects do not have a standardized spelling system, despite a literary history that goes back to the beginning of the 19th century. As a consequence, writers often use their own spelling systems, more or less based on German and often with some specifically French characters. But none of these systems can be seen as fully canonical. In this paper, we present the findings of an analysis of the spelling systems used in four different Alsatian datasets, including three newly transcribed lexicons, and describe how they differ by taking the phonetic transcriptions into account. We also detail experiments with a grapheme-to-phoneme (G2P) system trained on manually transcribed data and show that the combination of both spelling and phonetic variation presents specific challenges.

ChAnot: An Intelligent Annotation Tool for Indigenous and Highly Agglutinative Languages in Peru

Rodolfo Mercado, José Pereira, Marco Antonio Sobrevilla Cabezudo and Arturo Oncevay

Linguistic corpus annotation is one of the most important phases for addressing Natural Language Processing (NLP) tasks, as these methods are deeply involved with corpus-based techniques. However, meta-data annotation is a highly laborious manual task. A supportive alternative requires the use of computational tools. They are likely to simplify some of these operations, while can be adjusted appropriately to the needs of particular language features at the same time. Therefore, this paper presents ChAnot, a web-based annotation tool developed for Peruvian indigenous

and highly agglutinative languages, where Shipibo-Konibo was the case study. This new tool is able to support a diverse set of linguistic annotation tasks, such as morphological segmentation markup, POS-tag markup, among others. Also, it includes a suggestion engine based on historic and machine learning models, and a set of statistics about previous annotations.

The DLDP Survey on Digital Use and Usability of EU Regional and Minority Languages

Claudia Soria, Valeria Quochi and Irene Russo

The digital development of regional and minority languages requires careful planning to be effective and should be preceded by the identification of the current and actual extent to which those languages are used digitally, the type and frequency of their digital use, the opportunity for their use, and the main obstacles currently preventing it. This paper reports about the design, the results and the key findings of an exploratory survey launched by the Digital Language Diversity Project about the digital use and usability of regional and minority languages on digital media and devices. The aim of the survey - the first of this kind - was to investigate the real usage, needs and expectations of European minority language speakers regarding digital opportunities, with a strong focus on electronic communication. The survey is restricted to four languages (Basque, Breton, Karelian and Sardinian) at different stages of digital development, which offers a starting point to develop strategies for assessing digital vitality of these languages and overcoming specific difficulties such as, for instance, the lack of official data.

ASR for Documenting Acutely Under-Resourced Indigenous Languages

Robert Jimerson and Emily Prud'hommeaux

Despite its potential utility for facilitating the transcription of speech recordings, automatic speech recognition (ASR) has not been widely explored as a tool for documenting endangered languages. One obstacle to adopting ASR for this purpose is that the amount of data needed to build a reliable ASR system far exceeds what would typically be available in an endangered language. Languages with highly complex morphology present further data sparsity challenges. In this paper, we present a working ASR system for Seneca, an endangered indigenous language of North America, as a case study for the development of ASR for acutely low-resource languages in need of linguistic documentation. We explore methods of leveraging linguistic knowledge to improve the ASR language models for a polysynthetic language with few high-quality audio and text resources, and we propose a tool for using ASR output to bootstrap new data to iteratively improve the acoustic model. This work serves as a proof-of-concept for speech researchers interested helping field linguists and indigenous language community members engaged in the documentation and revitalization of endangered languages.

Session P57 - Opinion Mining / Sentiment Analysis (3)

11th May 2018, 11:45

Chair person: **Rodrigo Agerri** Poster Session

Building a Sentiment Corpus of Tweets in Brazilian Portuguese

Henrico Brum and Maria das Graças Volpe Nunes

The large amount of data available in social media, forums and websites motivates researches in several areas of Natural Language Processing, such as sentiment analysis. The popularity of the area due to its subjective and semantic characteristics motivates research on novel methods and approaches for classification. Hence, there is a high demand for datasets on different domains and different languages. This paper introduces TweetSentBR, a sentiment corpus for Brazilian Portuguese manually annotated with 15.000 sentences on TV show domain. The sentences were labeled in three classes (positive, neutral and negative) by seven annotators, following literature guidelines for ensuring reliability on the annotation. We also ran baseline experiments on polarity classification using six machine learning classifiers, reaching 80.38% on F-Measure in binary classification and 64.87% when including the neutral class. We also performed experiments in similar datasets for polarity classification task in comparison to this corpus.

'Aye' or 'No'? Speech-level Sentiment Analysis of Hansard UK Parliamentary Debate Transcripts

Gavin Abercrombie and Riza Batista-Navarro

Transcripts of UK parliamentary debates provide access to the opinions of politicians towards important topics, but due to the large quantity of textual data and the specialised language used, they are not straightforward for humans to process. We apply opinion mining methods to these transcripts to classify the sentiment polarity of speakers as being either positive or negative towards the motions proposed in the debates. We compare classification performance on a novel corpus using both manually annotated sentiment labels and labels derived from the speakers' votes ('aye' or 'no'). We introduce a two-step classification model, and evaluate the performance of both one- and two-step models, and the use of a range of textual and contextual features. Results suggest that textual features are more indicative of manually annotated class labels. Contextual metadata features however, boost performance are particularly indicative of vote

labels. Use of the two-step debate model results in performance gains and appears to capture some of the complexity of the debate format. Optimum performance on this data is achieved using all features to train a multi-layer neural network, indicating that such models may be most able to exploit the relationships between textual and contextual cues in parliamentary debate speeches.

Scalable Visualisation of Sentiment and Stance

Jon Chamberlain, Udo Kruschwitz and Orland Hoeber

Natural language processing systems have the ability to analyse not only the sentiment of human language, but also the stance of the speaker. Representing this information visually from unevenly distributed and potentially sparse datasets is challenging, in particular when trying to facilitate exploration and knowledge discovery. We present work on a novel visualisation approach for scalable visualisation of sentiment and stance and provide a language resource of e-government public engagement of 9,278 user comments with stance explicitly declared by the author.

NoReC: The Norwegian Review Corpus

Erik Velldal, Lilja Øvrelid, Eivind Alexander Bergem, Cathrine Stadsnes, Samia Touileb and Fredrik Jørgensen

This paper presents the Norwegian Review Corpus (NoReC), created for training and evaluating models for document-level sentiment analysis. The full-text reviews have been collected from major Norwegian news sources and cover a range of different domains, including literature, movies, video games, restaurants, music and theater, in addition to product reviews across a range of categories. Each review is labeled with a manually assigned score of 1–6, as provided by the rating of the original author. This first release of the corpus comprises more than 35,000 reviews. It is distributed using the CoNLL-U format, pre-processed using UDPipe, along with a rich set of metadata. The work reported in this paper forms part of the SANT initiative (Sentiment Analysis for Norwegian Text), a project seeking to provide open resources and tools for sentiment analysis and opinion mining for Norwegian.

SenSALDO: Creating a Sentiment Lexicon for Swedish

Jacobo Rouces, Nina Tahmasebi, Lars Borin and Stian Rødven Eide

The natural language processing subfield known as sentiment analysis or opinion mining has seen an explosive expansion over the last decade or so, and sentiment analysis has become a standard item in the NLP toolbox. Still, many theoretical and methodological questions remain unanswered and resource gaps unfilled. Most work on automated sentiment analysis has been done on English and a few other languages; for most written languages of the world, this tool is not available. This paper describes the development of an extensive sentiment lexicon for written (standard) Swedish. We investigate different methods for developing a sentiment lexicon for Swedish. We use an existing gold standard dataset for training and testing. For each word sense from the SALDO Swedish lexicon, we assign a real value sentiment score in the range [-1,1] and produce a sentiment label. We implement and evaluate three methods: a graph-based method that iterates over the SALDO structure, a method based on random paths over the SALDO structure and a corpus-driven method based on word embeddings. The resulting sense-disambiguated sentiment lexicon (SenSALDO) is an open source resource and freely available from Språkbanken, The Swedish Language Bank at the University of Gothenburg.

Corpus Building and Evaluation of Aspect-based Opinion Summaries from Tweets in Spanish

Daniel Peñaloza, Juanjosé Tenorio, Rodrigo López, Héctor Gomez, Arturo Oncevay and Marco Antonio Sobrevilla Cabezudo

This project involves the presentation and analysis of a corpus of Spanish extractive and abstractive summaries of opinions. The purpose of this work is to display a corpus of diverse summaries that could be used as a reference for academic research as we have not found one for the Spanish language as far as we know. We have analyzed the summaries based on the agreement between them as this shows how different they are written between each other and on aspect coverage and sentiment orientation as this proves the difference between the content that each summary tries to express. After the experimentation, we have found that even if each annotator uses a different expression to summarize a text, all of them contain similar messages. Furthermore, when writing, all of them prioritize on common aspects that are more representative of the corpus.

Application and Analysis of a Multi-layered Scheme for Irony on the Italian Twitter Corpus TWITTIRÒ

Alessandra Teresa Cignarella, Cristina Bosco, Viviana Patti and Mirko Lai

In this paper we describe the main issues emerged within the application of a multi-layered scheme for the fine-grained annotation of irony (Karoui et al., 2017) on an Italian Twitter corpus, i.e. TWITTIRÒ , which is composed of about 1,500 tweets with various provenance. A discussion is proposed about the limits and advantages of the application of the scheme to Italian messages, supported by an analysis of the outcome of the

annotation carried on by native Italian speakers in the development of the corpus. We present a quantitative and qualitative study both of the distribution of the labels for the different layers involved in the scheme which can shed some light on the process of human annotation for a validation of the annotation scheme on Italian irony-laden social media contents collected in the last years. This results in a novel gold standard for irony detection in Italian, enriched with fine-grained annotations, and in a language resource available to the community and exploitable in the cross- and multilingual perspective which characterizes the work that inspired this research.

Classifier-based Polarity Propagation in a WordNet

Jan Kocoń, Arkadiusz Janz and Maciej Piasecki

In this paper we present a novel approach to the construction of an extensive, sense-level sentiment lexicon built on the basis of a wordnet. The main aim of this work is to create a high-quality sentiment lexicon in a partially automated way. We propose a method called Classifier-based Polarity Propagation, which utilises a very rich set of wordnet-based features, to recognize and assign specific sentiment polarity values to wordnet senses. We have demonstrated that in comparison to the existing rule-base solutions using specific, narrow set of semantic relations, our method allows for the construction of a more reliable sentiment lexicon, starting with the same seed of annotated synsets.

Session P58 - Sign Language

11th May 2018, 11:45

Chair person: **Thomas Hanke** Poster Session

SMILE Swiss German Sign Language Dataset

Sarah Ebling, Necati Cihan Camgöz, Penny Boyes Braem, Katja Tissi, Sandra Sidler-Miserez, Stephanie Stoll, Simon Hadfield, Tobias Haug, Richard Bowden, Sandrine Tornay, Marzieh Razavi and Mathew Magimai-Doss

Sign language recognition (SLR) involves identifying the form and meaning of isolated signs or sequences of signs. To our knowledge, the combination of SLR and sign language assessment is novel. The goal of an ongoing three-year project in Switzerland is to pioneer an assessment system for lexical signs of Swiss German Sign Language (Deutschschweizerische Gebärdensprache, DSGS) that relies on SLR. The assessment system aims to give adult L2 learners of DSGS feedback on the correctness of the manual parameters (handshape, hand position, location, and movement) of isolated signs they produce. In its initial version, the system will include automatic feedback for a subset of a DSGS vocabulary production test consisting of 100 lexical items. To provide the SLR component of the assessment system with sufficient training samples, a large-scale dataset containing videotaped repeated productions of the 100 items of the vocabulary test with associated transcriptions and annotations was created, consisting of data from 11 adult L1 signers and 19 adult L2 learners of DSGS. This paper introduces the dataset, which will be made available to the research community.

IPSL: A Database of Iconicity Patterns in Sign Languages. Creation and Use

Vadim Kimmelman, Anna Klezovich and George Moroz

We created the first large-scale database of signs annotated according to various parameters of iconicity. The signs represent concrete concepts in seven semantic fields in nineteen sign languages; 1542 signs in total. Each sign was annotated with respect to the type of form-image association, the presence of iconic location and movement, personification, and with respect to whether the sign depicts a salient part of the concept. We also created a website: https://sl-iconicity.shinyapps.io/iconicity_patterns/ with several visualization tools to represent the data from the database. It is possible to visualize iconic properties of separate concepts or iconic properties of semantic fields on the map of the world, and to build graphs representing iconic patterns for selected semantic fields. A preliminary analysis of the data shows that iconicity patterns vary across semantic fields and across languages. The database and the website can be used to further study a variety of theoretical questions related to iconicity in sign languages.

Sign Languages and the Online World Online Dictionaries & Lexicostatistics

Shi Yu, Carlo Geraci and Natasha Abner

Several online dictionaries documenting the lexicon of a variety of sign languages (SLs) are now available. These are rich resources for comparative studies, but there are methodological issues that must be addressed regarding how these resources are used for research purposes. We created a web-based tool for annotating the articulatory features of signs (handshape, location, movement and orientation). Videos from online dictionaries may be embedded in the tool, providing a mechanism for large-scale theoretically-informed sign language annotation. Annotations are saved in a spreadsheet format ready for quantitative and qualitative analyses. Here, we provide proof of concept for the utility of this tool in linguistic analysis. We used the SL adaptation of the Swadesh list (Woodward, 2000) and applied lexicostatistic and phylogenetic methods to a sample of 23 SLs coded using the web-based tool; supplementary historic information was gathered from the Ethnologue of World Languages and other online sources. We report results from the comparison of all articulatory features for

four Asian SLs (Chinese, Hong Kong, Taiwanese and Japanese SLs) and from the comparison of handshapes on the entire 23 language sample. Handshape analysis of the entire sample clusters all Asian SLs together, separated from the European, American, and Brazilian SLs in the sample, as historically expected. Within the Asian SL cluster, analyses also show, for example, marginal relatedness between Chinese and Hong Kong SLs.

Elicitation protocol and material for a corpus of long prepared monologues in Sign Language

Michael Filhol and Mohamed Nassime Hadjadj

In this paper, we address collection of prepared Sign Language discourse, as opposed to spontaneous signing. Specifically, we aim at collecting long discourse, which creates problems explained in the paper. Being oral and visual languages, they cannot easily be produced while reading notes without distorting the data, and eliciting long discourse without influencing the production order is not trivial. For the moment, corpora contain either short productions, data distortion or disfluencies. We propose a protocol and two tasks with their elicitation material to allow cleaner long-discourse data, and evaluate the result of a recent test with LSF informants.

Deep JSLC: A Multimodal Corpus Collection for Data-driven Generation of Japanese Sign Language Expressions

Heike Brock and Kazuhiro Nakadai

The three-dimensional visualization of spoken or written information in Sign Language (SL) is considered a potential tool for better inclusion of deaf or hard of hearing individuals with low literacy skills. However, conventional technologies for such CG-supported data display are not able to depict all relevant features of a natural signing sequence such as facial expression, spatial references or inter-sign movement, leading to poor acceptance amongst speakers of sign language. The deployment of fully data-driven, deep sequence generation models that proved themselves powerful in speech and text applications might overcome this lack of naturalness. Therefore, we collected a corpus of continuous sentence utterances in Japanese Sign Language (JSL) applicable to the learning of deep neural network models. The presented corpus contains multimodal content information of high resolution motion capture data, video data and both visual and gloss-like mark up annotations obtained with the support of fluent JSL signers. Furthermore, all annotations were encoded under three different encoding schemes with respect to directions, intonation and non-manual information. Currently, the corpus is employed to learn first sequence-to-sequence networks where it shows the ability to train relevant language features.

Modeling French Sign Language: a proposal for a semantically compositional system

Mohamed nassime Hadjadj, Michael Filhol and Annelies Braffort

The recognition of French Sign Language (LSF) as a natural language in 2005 created an important need for the development of tools to make information accessible to the deaf public. With this prospect, the goal of this article is to propose a linguistic approach aimed at modeling the French sign language. We first present the models proposed in computer science to formalize the sign language (SL). We also show the difficulty of applying the grammars originally designed for spoken languages to model SL. In a second step, we propose an approach allowing to take into account the linguistic properties of the SL while respecting the constraints of a modelisation process. By studying the links between semantic functions and their observed forms in Corpus, we have identified several production rules that govern the functioning of the LSF. We finally present the rule functioning as a system capable of modeling an entire utterance in French sign language.

Session P59 - Speech Resource/Database (2)

11th May 2018, 11:45

Chair person: **Christoph Draxler** Poster Session

Construction of the Corpus of Everyday Japanese Conversation: An Interim Report

Hanae Koiso, Yasuharu Den, Yuriko Iseki, Wakako Kashino, Yoshiko Kawabata, Ken'ya Nishikawa, Yayoi Tanaka and Yasuyuki Usuda

In 2016, we launched a new corpus project in which we are building a large-scale corpus of everyday Japanese conversation in a balanced manner, aiming at exploring characteristics of conversations in contemporary Japanese through multiple approaches. The corpus targets various kinds of naturally occurring conversations in daily situations, such as conversations during dinner with the family at home, meetings with colleagues at work, and conversations while driving. In this paper, we first introduce an overview of the corpus, including corpus size, conversation variations, recording methods, structure of the corpus, and annotations to be included in the corpus. Next, we report on the current stage of the development of the corpus and legal and ethical issues discussed so far. Then we present some results of the preliminary evaluation of the data being collected. We focus on whether or not the 94 hours of conversations collected so far vary in a balanced manner by reference to the survey results of everyday conversational behavior that we conducted previously

to build an empirical foundation for the corpus design. We will publish the whole corpus in 2022, consisting of more than 200 hours of recordings.

Carcinologic Speech Severity Index Project: A Database of Speech Disorder Productions to Assess Quality of Life Related to Speech After Cancer

Corine Astésano, Mathieu Balaguer, Jérôme Farinas, Corinne Fredouille, Pascal Gaillard, Alain Ghio, Imed Laaridh, Muriel Lalain, Benoît Lepage, Julie Mauclair, Olivier Nocaudie, Julien Pinquier, Oriol Pont, Gilles Pouchoulin, Michèle Puech, Danièle Robert, Etienne Sicard and Virginie Woisard

Within the framework of the Carcinologic Speech Severity Index (C2SI) InCA Project, we collected a large database of French speech recordings aiming at validating Disorder Severity Indexes. Such a database will be useful for measuring the impact of oral and pharyngeal cavity cancer on speech production. That will permit to assess patients' Quality of Life after treatment. The database is composed of audio recordings from 135 speakers and associated metadata. Several intelligibility and comprehensibility levels of speech functions have been evaluated. Acoustics and Prosody have been assessed. Perceptual evaluation rates from both naive and expert juries are being produced. Automatic analyzes are being carried out. That will provide to speech therapists objective tools to take into account the intelligibility and comprehensibility of patients which received cancer treatment (surgery and/or radiotherapy and/or chemotherapy). The aim of this paper is to justify the need of this corpus and his data collection. This corpus will be available to the scientific community through the GIS Parolotheque.

Parallel Corpora in Mboshi (Bantu C25, Congo-Brazzaville)

Annie Rialland, Martine Adda-Decker, Guy-Noël Kouarata, Gilles Adda, Laurent Besacier, Lori Lamel, Elodie Gauthier, Pierre Godard and Jamison Cooper-Leavitt

This article presents multimodal and parallel data collections in Mboshi, as part of the French-German BULB project. It aims at supporting documentation and providing digital resources for less resourced languages with the help of speech and language-based technology. The data collection specifications thus have to meet both field linguists' and computer scientists' requirements, which are large corpora for the latter and linguistically dense data for the former. Beyond speech, the collection comprises pictures and videos documenting social practices, agriculture, wildlife and plants. Visual supports aimed at encouraging people to comment on objects which are meaningful in their daily lives. Speech recordings are composed of the original speech in Mboshi, a respoken version and a translated version to French. These three speech streams remain time-aligned thanks to LIG-AIKUMA, which adds new features to a previous AIKUMA application. The speech corpus includes read material (5k sentences, Bible), verb conjugations and a large part of spontaneous speech (conversations, picture descriptions) resulting in over 50 hours of Mboshi speech, of which 20 hours are already respoken and orally translated to French. These parallel oral data are intended for linguistic documentation (tonology, phonology...) and automatic processing (corpus annotation, alignment between Mboshi speech and French translations).

A Multimodal Corpus of Expert Gaze and Behavior during Phonetic Segmentation Tasks

Arif Khan, Ingmar Steiner, Yusuke Sugano, Andreas Bulling and Ross Macdonald

Phonetic segmentation is the process of splitting speech into distinct phonetic units. Human experts routinely perform this task manually by analyzing auditory and visual cues using analysis software, which is an extremely time-consuming process. Methods exist for automatic segmentation, but these are not always accurate enough. In order to improve automatic segmentation, we need to model it as close to the manual segmentation as possible. This corpus is an effort to capture the human segmentation behavior by recording experts performing a segmentation task. We believe that this data will enable us to highlight the important aspects of manual segmentation, which can be used in automatic segmentation to improve its accuracy.

Statistical Analysis of Missing Translation in Simultaneous Interpretation Using A Large-scale Bilingual Speech Corpus

Zhongxi Cai, Koichiro Ryu and Shigeki Matsubara

This paper describes statistical analyses of missing translations in simultaneous interpretations. Eighty-eight lectures from English-to-Japanese interpretation data from a large-scale bilingual speech corpus were used for the analyses. Word-level alignment was provided manually, and English words without corresponding Japanese words were considered missing translations. The English lectures contained 46,568 content words, 33.1\% of which were missing in the translation. We analyzed the relationship between missing translations and various factors, including the speech rate of the source language, delay of interpretation, part-of-speech, and depth in the syntactic structure of the source language. The analyses revealed that the proportion of missing translations is high when the speech rate is high and delay is large. We also found

that a high proportion of adverbs were missed in the translations, and that words at deeper positions in the syntactic structure were more likely to be missed.

SynPaFlex-Corpus: An Expressive French Audiobooks Corpus dedicated to expressive speech synthesis.

Aghilas SINI, Damien Lolive, Gaëlle Vidal, Marie Tahon and Élisabeth Delais-Roussarie

This paper presents an expressive French audiobooks corpus containing eighty seven hours of good audio quality speech, recorded by a single amateur speaker reading audiobooks of different literary genres. This corpus departs from existing corpora collected from audiobooks since they usually provide a few hours of mono-genre and multi-speaker speech. The motivation for setting up such a corpus is to explore expressiveness from different perspectives, such as discourse styles, prosody, and pronunciation, and using different levels of analysis (syllable, prosodic and lexical words, prosodic and syntactic phrases, utterance or paragraph). This will allow developing models to better control expressiveness in speech synthesis, and to adapt pronunciation and prosody to specific discourse settings (changes in discourse perspectives, indirect vs. direct styles, etc.). To this end, the corpus has been annotated automatically and provides information as phone labels, phone boundaries, syllables, words or morpho-syntactic tagging. Moreover, a significant part of the corpus has also been annotated manually to encode direct/indirect speech information and emotional content. The corpus is already usable for studies on prosody and TTS purposes and is available to the community.

Increasing the Accessibility of Time-Aligned Speech Corpora with Spokes Mix

Piotr Pęzik

Spokes Mix is an online service providing access to a number of spoken corpora of Polish, including three newly released time-aligned collections of manually transcribed spoken-conversational data. The purpose of this service is two-fold. Firstly, it functions as a programmatic interface to a number of unique collections of conversational Polish and potentially also spoken corpora of other languages, exposing their full content with complete metadata and annotations. Equally important, however, is its second function of increasing the general accessibility of these resources for research on spoken and conversational language by providing a centralized, easy-to-use corpus query engine with a responsive web-based user interface.

The MonPaGe_HA Database for the Documentation of Spoken French Throughout Adulthood

Cécile Fougeron, Veronique Delvaux, Lucie Ménard and Marina Laganaro

Our knowledge of life-span changes in the speech of adults is quite sparse. Existing reports are mainly based on English speakers and few studies have compared more than two extreme age groups. The present paper describes the recently constituted MonPaGe_HealthyAdults database of spoken French including 405 male and female speakers aged from 20 to 93 years old. This database aims at documenting speech throughout adulthood and at building a set of reference values for healthy speakers to be used in clinical assessment of speech. The database is built on five age groups ([20-39], [40-49], [50-59], [60-74], [75+]) and includes 4 regiolects. Speakers have been recorded on a variety of linguistic material and speech tasks in order to cover multiple speech dimensions for each speaker. These cross-sectional data form one of the largest French database available for observing typical changes in the speech of adults as a function of age, and especially in older adults.

Bringing Order to Chaos: A Non-Sequential Approach for Browsing Large Sets of Found Audio Data

Per Fallgren, Zofia Malisz and Jens Edlund

We present a novel and general approach for fast and efficient non-sequential browsing of sound in large archives that we know little or nothing about, e.g. so called found data – data not recorded with the specific purpose to be analysed or used as training data. Our main motivation is to address some of the problems speech and speech technology researchers see when they try to capitalise on the huge quantities of speech data that reside in public archives. Our method is a combination of audio browsing through massively multi-object sound environments and a well-known unsupervised dimensionality reduction algorithm (SOM). We test the process chain on four data sets of different nature (speech, speech and music, farm animals, and farm animals mixed with farm sounds). The methods are shown to combine well, resulting in rapid and readily interpretable observations. Finally, our initial results are demonstrated in prototype software which is freely available.

CoLoSS: Cognitive Load Corpus with Speech and Performance Data from a Symbol-Digit Dual-Task

Robert Herms, Maria Wirzberger, Maximilian Eibl and Günter Daniel Rey

In this paper, a new corpus named CoLoSS (Cognitive Load by Speech and performance data in a Symbol-digit dual-task) is

presented, which contains speech under cognitive load recorded in a learning task scenario. In order to obtain a reference for cognitive load, a dual-task approach was applied, including a visual-motor primary task that required subjects to learn abstract symbol combinations and an auditory-verbal secondary task to measure the load imposed by the primary task. We report the methodology of collecting the speech recordings, constructing the corpus and describe the properties of the data. Finally, effects of cognitive load on prosodic as well as voice quality features are investigated in conjunction with the corpus. In its current version, the corpus is available to the scientific community, e.g., for exploring the influence of cognitive load on speech or conducting experiments for speech-based cognitive load recognition.

VAST: A Corpus of Video Annotation for Speech Technologies

Jennifer Tracey and Stephanie Strassel

The Video Annotation for Speech Technologies (VAST) corpus contains approximately 2900 hours of video data collected and labeled to support the development of speech technologies such as speech activity detection, language identification, speaker identification, and speech recognition. The bulk of the data comes from amateur video content harvested from the web. Collection was designed to ensure that the videos cover a diverse range of communication domains, data sources and video resolutions and to include three primary languages (English, Mandarin Chinese and Arabic) plus supplemental data in 7 additional languages/dialects to support language recognition research. Portions of the collected data were annotated for speech activity, speaker identity, speaker sex, language identification, diarization, and transcription. A description of the data collection and each of the annotation types is presented in this paper. The corpus represents a challenging data set for language technology development due to the informal nature of the majority of the data, as well as the variety of languages, noise conditions, topics, and speakers present in the collection.

Session O37 - Anaphora & Coreference

11th May 2018, 14:45

Chair person: **Claire Bonial** Oral Session

A Gold Anaphora Annotation Layer on an Eye Movement Corpus

Olga Seminck and Pascal Amsili

Anaphora resolution is a complex process in which multiple linguistic factors play a role, and this is witnessed by a large psycholinguistic literature. This literature is based on experiments with hand-constructed items, which have the advantage to filter influences outside the scope of the study, but, as a downside, make the experimental data artificial. Our goal is to provide a first resource allowing to study human anaphora resolution on natural data. We annotated anaphorical pronouns in the Dundee Corpus: a corpus of 50k words coming from newspaper articles read by humans of whom all eye movements were recorded. We identified all anaphoric pronouns - in opposition to non-referential, cataphoric and deictic uses - and identified the closest antecedent for each of them. Both the identification of the anaphoricity and the antecedents of the pronouns showed a high inter-annotator agreement. We used our resource to model reading time of pronouns to study simultaneously various factors of influence on anaphora resolution. Whereas the influence of the anaphoric relation on the reading time of the pronoun is subtle, psycholinguistic findings from settings using experimental items were confirmed. In this way our resource provides a new means to study anaphora.

Annotating Zero Anaphora for Question Answering

Yoshihiko Asao, Ryu Iida and Kentaro Torisawa

We constructed a large annotated dataset of zero pronouns that correspond to adjuncts marked by -de (translated to English as 'in', 'at', 'by' or 'with') in Japanese. Adjunct zero anaphora resolution plays an important role in extracting information such as location and means from a text. To our knowledge, however, there have been no large-scale dataset covering them. In this paper, focusing on the application of zero anaphora resolution to question answering (QA), we proposed two annotation schemes. The first scheme was designed to efficiently collect zero anaphora instances that are useful in QA. Instead of directly annotating zero anaphora, annotators evaluated QA instances whose correctness hinges on zero anaphora resolution. Over 20,000 instances of zero anaphora were collected with this scheme. We trained a multi-column convolutional neural network with the annotated data, achieving an average precision of 0.519 in predicting the correctness of QA instances of the same type. In the second scheme, zero anaphora is annotated in a more direct manner. A model trained with the results of the second annotation scheme performed better than the first scheme in identifying zero anaphora for sentences randomly sampled from a corpus, suggesting a tradeoff between application-specific and general-purpose annotation schemes.

Building Named Entity Recognition Taggers via Parallel Corpora

Rodrigo Agerri, Yiling Chung, Itziar Aldabe, Nora Aranberri, Gorka Labaka and German Rigau

The lack of hand curated data is a major impediment to developing statistical semantic processors for many of the world languages. Our paper aims to bridge this gap by leveraging existing annotations and semantic processors from multiple source languages by projecting their annotations via the statistical word alignments traditionally used in Machine Translation. Taking the Named Entity Recognition (NER) task as a use case, this work presents a method to automatically induce Named Entity annotated data using parallel corpora without any manual intervention. The projected annotations can then be used to automatically generate semantic processors for the target language helping to overcome the lack of training data for a given language. The experiments are focused on 4 languages: German, English, Spanish and Italian, and our empirical evaluation results show that our method obtains competitive results when compared with models trained on gold-standard, albeit out-of-domain, data. The results point out that our projection algorithm is effective to transport NER annotations across languages thus providing a fully automatic method to obtain NER taggers for as many as the number of languages aligned in parallel corpora. Every resource generated (training data, manually annotated test set and NER models) is made publicly available for its use and to facilitate reproducibility of results.

Cross-Document, Cross-Language Event Coreference Annotation Using Event Hoppers

Zhiyi Song, Ann Bies, Justin Mott, Xuansong Li, Stephanie Strassel and Christopher Caruso

We discuss the development and implementation of an approach for cross-document, cross-lingual event coreference for the DEFT Rich Entities, Relations and Events (Rich ERE) annotation task. Rich ERE defined the notion of event hoppers to enable intuitive within-document coreference for the DEFT event ontology, and the expansion of coreference to cross-document, cross-lingual event mentions relies crucially on this same construct. We created new annotation guidelines, data processes and user interfaces to enable annotation of 505 documents in three languages selected from data already labeled for Rich ERE, yielding 389 cross-document event hoppers. We discuss the data creation process and the central role of event hoppers in making cross-document, cross-lingual coreference decisions. We present the challenges encountered during annotation along with three directions for future work.

Session O38 - Corpus for Document Classification

11th May 2018, 14:45

Chair person: **Franciska de Jong** Oral Session

TAP-DLND 1.0 : A Corpus for Document Level Novelty Detection

Tirthankar Ghosal, Amitra Salam, Swati Tiwary, Asif Ekbal and Pushpak Bhattacharyya

Detecting novelty of an entire document is an Artificial Intelligence (AI) frontier problem. This has immense importance in widespread Natural Language Processing (NLP) applications ranging from extractive text document summarization to tracking development of news events to predicting impact of scholarly articles. Although a very relevant problem in the present context of exponential data duplication, we are unaware of any document level dataset that correctly addresses the evaluation of automatic novelty detection techniques in a classification framework. To bridge this relative gap, here in this work, we present a resource for benchmarking the techniques for document level novelty detection. We create the resource via topic-specific crawling of news documents across several domains in a periodic manner. We release the annotated corpus with necessary statistics and show its use with a developed system for the problem in concern.

A Corpus for Multilingual Document Classification in Eight Languages

Holger Schwenk and Xian Li

Cross-lingual document classification aims at training a document classifier on resources in one language and transferring it to a different language without any additional resources. Several approaches have been proposed in the literature and the current best practice is to evaluate them on a subset of the Reuters Corpus Volume 2. However, this subset covers only few languages (English, German, French and Spanish) and almost all published works focus on the the transfer between English and German. In addition, we have observed that the class prior distributions differ significantly between the languages. We argue that this complicates the evaluation of the multilinguality. In this paper, we propose a new subset of the Reuters corpus with balanced class priors for eight languages. By adding Italian, Russian, Japanese and Chinese, we cover languages which are very different with respect to syntax, morphology, etc. We provide strong baselines for all language transfer directions using multilingual word

and sentence embeddings respectively. Our goal is to offer a freely available framework to evaluate cross-lingual document classification, and we hope to foster by these means, research in this important area.

Analyzing Citation-Distance Networks for Evaluating Publication Impact

Drahomira Herrmannova, Petr Knoth and Robert Patton

Studying citation patterns of scholarly articles has been of interest to many researchers from various disciplines. While the relationship of citations and scientific impact has been widely studied in the literature, in this paper we develop the idea of analyzing the semantic distance of scholarly articles in a citation network (citation-distance network) to uncover patterns that reflect scientific impact. More specifically, we compare two types of publications in terms of their citation-distance patterns, seminal publications and literature reviews, and focus on their referencing patterns as well as on publications which cite them. We show that seminal publications are associated with a larger semantic distance, measured using the content of the articles, between their references and the citing publications, while literature reviews tend to cite publications from a wider range of topics. Our motivation is to understand and utilize this information to create new research evaluation metrics which would better reflect scientific impact.

Annotating Educational Questions for Student Response Analysis

Andreea Godea and Rodney Nielsen

Questions play an important role in the educational domain, representing the main form of interaction between instructors and students. In this paper, we introduce the first taxonomy and annotated educational corpus of questions that aims to help with the analysis of student responses. The dataset can be employed in approaches that classify questions based on the expected answer types. This can be an important component in applications that require prior knowledge about the desired answer to a given question, such as educational and question answering systems. To demonstrate the applicability and the effectiveness of the data within approaches to classify questions based on expected answer types, we performed extensive experiments on our dataset using a neural network with word embeddings as features. The approach achieved a weighted F1-score of 0.511, overcoming the baseline by 12%. This demonstrates that our corpus can be effectively integrated in simple approaches that classify questions based on the response type.

Session O39 - Knowledge Discovery & Evaluation (2)

11th May 2018, 14:45

Chair person: **Yuji Matsumoto** Oral Session

Incorporating Global Contexts into Sentence Embedding for Relational Extraction at the Paragraph Level with Distant Supervision

Eun-kyung Kim and KEY-SUN CHOI

The increased demand for structured knowledge has created considerable interest in relation extraction (RE) from large collections of documents. In particular, distant supervision can be used for RE without manual annotation costs. Nevertheless, this paradigm only extracts relations from individual sentences that contain two target entities. This paper explores the incorporation of global contexts derived from paragraph-into-sentence embedding as a means of compensating for the shortage of training data in distantly supervised RE. Experiments on RE from Korean Wikipedia show that the presented approach can learn an exact RE from sentences (including grammatically incoherent sentences) without syntactic parsing.

MCScript: A Novel Dataset for Assessing Machine Comprehension Using Script Knowledge

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater and Manfred Pinkal

We introduce a large dataset of narrative texts and questions about these texts, intended to be used in a machine comprehension task that requires reasoning using commonsense knowledge. Our dataset complements similar datasets in that we focus on stories about everyday activities, such as going to the movies or working in the garden, and that the questions require commonsense knowledge, or more specifically, script knowledge, to be answered. We show that our mode of data collection via crowdsourcing results in a substantial amount of such inference questions. The dataset forms the basis of a shared task on commonsense and script knowledge organized at SemEval 2018 and provides challenging test cases for the broader natural language understanding community.

A Neural Network Based Model for Loanword Identification in Uyghur

Chenggang Mi, Yating Yang, Lei Wang, Xi Zhou and Tonghai Jiang

Lexical borrowing happens in almost all languages. To obtain more bilingual knowledge from monolingual corpora, we propose

a neural network based loanword identification model for Uyghur. We build our model on a bidirectional LSTM - CNN framework, which can capture past and future information effectively and learn both word level and character level features from training data automatically. To overcome data sparsity that exists in model training, we also suggest three additional features , such as hybrid language model feature, pronunciation similarity feature and part-of-speech tagging feature to further improve the performance of our proposed approach. We conduct experiments on Chinese, Arabic and Russian loanword detection in Uyghur. Experimental results show that our proposed method outperforms several baseline models.

Revisiting Distant Supervision for Relation Extraction

Tingsong Jiang, Jing Liu, Chin-Yew Lin and Zhifang Sui

Distant supervision has been widely used in the task of relation extraction (RE). However, when we carefully examine the experimental settings of previous work, we find two issues: (i) The compared models were trained on different training datasets. (ii) The existing testing data contains noise and bias issues. These issues may affect the conclusions in previous work. In this paper, our primary aim is to re-examine the distant supervision-based approaches under the experimental settings without the above issues. We approach this by training models on the same dataset and creating a new testing dataset annotated by the workers on Amzaon Mechanical Turk. We draw new conclusions based on the new testing dataset. The new testing data can be obtained from http://aka.ms/relationie.

Session O40 - Multimodal & Written Corpora & Tools

11th May 2018, 14:45

Chair person: **Nancy Ide** Oral Session

Incorporating Contextual Information for Language-Independent, Dynamic Disambiguation Tasks

Tobias Staron, Özge Alacam and Wolfgang Menzel

Humans resolve various kinds of linguistic ambiguities by exploiting available external evidence that has been acquired from modalities besides the linguistic one. This behavior can be observed for several languages, for English or German for example. In contrast, most natural language processing systems, parsers for example, rely on linguistic information only without taking further knowledge into account. While those systems are expected to correctly handle syntactically unambiguous cases, they cannot resolve syntactic ambiguities reliably. This paper hypothesizes that parsers would be able to find non-canonical interpretations of ambiguous sentences, if they exploited external, contextual information. The proposed multi-modal system, which combines data-driven and grammar-based approaches, confirmed this hypothesis in experiments on syntactically ambiguous sentences. This work focuses on the scarcely investigated relative clause attachment ambiguity instead of prepositional phrase attachment ambiguities, which are already well known in the literature. Experiments were conducted for English, German and Turkish and dynamic, i. e. highly dissimilar, contexts.

Overcoming the Long Tail Problem: A Case Study on CO2-Footprint Estimation of Recipes using Information Retrieval

Melanie Geiger and Martin Braschler

We propose approaches that use information retrieval methods for the automatic calculation of CO2-footprints of cooking recipes. A particular challenge is the "long tail problem" that arises with the large diversity of possible ingredients. The proposed approaches are generalizable to other use cases in which a numerical value for semi-structured items has to be calculated, for example, the calculation of the insurance value of a property based on a real estate listing. Our first approach, ingredient matching, calculates the CO2-footprint based on the ingredient descriptions that are matched to food products in a language resource and therefore suffers from the long tail problem. On the other hand, our second approach directly uses the recipe to estimate the CO2-value based on its closest neighbor using an adapted version of the BM25 weighting scheme. Furthermore, we combine these two approaches in order to achieve a more reliable estimate. Our experiments show that the automatically calculated CO2-value estimates lie within an acceptable range compared to the manually calculated values. Therefore, the costs of the calculation of the CO2-footprints can be reduced dramatically by using the automatic approaches. This helps to make the information available to a large audience in order to increase the awareness and transparency of the environmental impact of food consumption.

Comparison of Pun Detection Methods Using Japanese Pun Corpus

Motoki Yatsu and Kenji Araki

A sampling survey of typology and component ratio analysis in Japanese puns revealed that the type of Japanese pun that had the largest proportion was a pun type with two sound sequences, whose consonants are phonetically close to each other in the same sentence which includes the pun. Based on this finding, we

constructed rules to detect pairs of phonetically similar sequences as features for a supervised machine learning classifier. Using these features in addition to Bag-of-Words features, an evaluation experiment confirmed the effectiveness of adding the rule-based features to the baseline.

A vision-grounded dataset for predicting typical locations for verbs

Nelson Mukuze, Anna Rohrbach, Vera Demberg and Bernt Schiele

Information about the location of an action is often implicit in text, as humans can infer it based on common sense knowledge. Today's NLP systems however struggle with inferring information that goes beyond what is explicit in text. Selectional preference estimation based on large amounts of data provides a way to infer prototypical role llers, but text-based systems tend to underestimate the probability of the most typical role llers. We here present a new dataset containing thematic t judgments for 2,000 verb/location pairs. This dataset can be used for evaluating text-based, vision-based or multimodal inference systems for the typicality of an event's location. We additionally provide three thematic t baselines for this dataset: a state-of-the-art neural networks based thematic t model learned from linguistic data, a model estimating typical locations based on the MSCOCO dataset and a simple combination of the systems.

Session P60 - Corpus Creation, Annotation, Use (2)

11th May 2018, 14:45

Chair person: **Beatrice Daille** Poster Session

Edit me: A Corpus and a Framework for Understanding Natural Language Image Editing

Ramesh Manuvinakurike, Jacqueline Brixey, Trung Bui, Walter Chang, Doo Soon Kim, Ron Artstein and Kallirroi Georgila

This paper introduces the task of interacting with an image editing program through natural language. We present a corpus of image edit requests which were elicited for real world images, and an annotation framework for understanding such natural language instructions and mapping them to actionable computer commands. Finally, we evaluate crowd-sourced annotation as a means of efficiently creating a sizable corpus at a reasonable cost.

Enriching a Lexicon of Discourse Connectives with Corpus-based Data

Anna Feltracco, Elisabetta Jezek and Bernardo Magnini

We present the results of the effort of enriching the pre-existing resource LICO, a Lexicon of Italian COnnectives retrieved from lexicographic sources (Feltracco et al., 2016), with real corpus data for connectives marking contrast relations in text. The motivation beyond our effort is that connectives can only be interpreted when they appear in context, that is, in a relation between the two fragments of text that constitute the two arguments of the relation. In this perspective, adding corpus examples annotated with connectives and arguments for the relation allows us to both extend the resource and validate the lexicon. In order to retrieve good corpus examples, we take advantage of the existing Contrast-Ita Bank (Feltracco et al., 2017), a corpus of news annotated with explicit and implicit discourse contrast relations for Italian according to the annotation scheme proposed in the Penn Discourse Tree Bank (PDTB) guidelines (Prasad et al., 2007). We also use an extended -non contrast annotated- version of the same corpus and documents from Wikipedia. The resulting resource represents a valuable tool for both linguistic analyses of discourse relations and the training of a classifier for NLP applications.

SimPA: A Sentence-Level Simplification Corpus for the Public Administration Domain

Carolina Scarton, Gustavo Paetzold and Lucia Specia

We present a sentence-level simplification corpus with content from the Public Administration (PA) domain. The corpus contains 1,100 original sentences with manual simplifications collected through a two-stage process. Firstly, annotators were asked to simplify only words and phrases (lexical simplification). Each sentence was simplified by three annotators. Secondly, one lexically simplified version of each original sentence was further simplified at the syntactic level. In its current version there are 3,300 lexically simplified sentences plus 1,100 syntactically simplified sentences. The corpus will be used for evaluation of text simplification approaches in the scope of the EU H2020 SIMPATICO project - which focuses on accessibility of e-services in the PA domain - and beyond. The main advantage of this corpus is that lexical and syntactic simplifications can be analysed and used in isolation. The lexically simplified corpus is also multi-reference (three different simplifications per original sentence). This is an ongoing effort and our final aim is to collect manual simplifications for the entire set of original sentences, with over 10K sentences.

The brWaC Corpus: A New Open Resource for Brazilian Portuguese

Jorge Alberto Wagner Filho, Rodrigo Wilkens, Marco Idiart and Aline Villavicencio

In this work, we present the construction process of a large Web corpus for Brazilian Portuguese, aiming to achieve a size comparable to the state of the art in other languages. We also discuss our updated sentence-level approach for the strict removal of duplicated content. Following the pipeline methodology, more than 60 million pages were crawled and filtered, with 3.5 million being selected. The obtained multi-domain corpus, named brWaC, is composed by 2.7 billion tokens, and has been annotated with tagging and parsing information. The incidence of non-unique long sentences, an indication of replicated content, which reaches 9% in other Web corpora, was reduced to only 0.5%. Domain diversity was also maximized, with 120,000 different websites contributing content. We are making our new resource freely available for the research community, both for querying and downloading, in the expectation of aiding in new advances for the processing of Brazilian Portuguese.

Czech Text Document Corpus v 2.0

Pavel Kral and Ladislav Lenc

This paper introduces "Czech Text Document Corpus v 2.0", a collection of text documents for automatic document classification in Czech language. It is composed of the text documents provided by the Czech News Agency and is freely available for research purposes at http://ctdc.kiv.zcu.cz/. This corpus was created in order to facilitate a straightforward comparison of the document classification approaches on Czech data. It is particularly dedicated to evaluation of multi-label document classification approaches, because one document is usually labelled with more than one label. Besides the information about the document classes, the corpus is also annotated at the morphological layer. This paper further shows the results of selected state-of-the-art methods on this corpus to offer the possibility of an easy comparison with these approaches.

Corpora of Typical Sentences

Lydia Müller, Uwe Quasthoff and Maciej Sumalvico

Typical sentences of characteristic syntactic structures can be used for language understanding tasks like finding typical slotfiller for verbs. The paper describes the selection of such typical sentences representing usually about 5% of the original corpus. The sentences are selected by the frequency of the corresponding POS tag sequence together with an entropy theshold, and the selection method is shown to work language independently. Entropy measuring the distribution of words in a given position turns out to identify larger sets of near-duplicate sentences, not considered typical. A statistical comparison of those subcorpora with the underlying corpus shows the intended shorter sentence length, but also a decrease of word frequencies for function words associated to more complex sentences.

The German Reference Corpus DeReKo: New Developments – New Opportunities

Marc Kupietz, Harald Lüngen, Pawel Kamocki and Andreas Witt

This paper discusses current trends in DeReKo, the German Reference Corpus, concerning legal issues around the recent German copyright reform with positive implications for corpus building and corpus linguistics in general, recent corpus extensions in the genres of popular magazines, journals, historical texts, and web-based football reports. Besides, DeReKo is finally accessible via the new corpus research platform KorAP, offering registered users several news features in comparison with its predecessor COSMAS II.

Risamálheild: A Very Large Icelandic Text Corpus

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson and Jon Gudnason

We present Risamálheild, the Icelandic Gigaword Corpus (IGC), a corpus containing more than one billion running words from mostly contemporary texts. The work was carried out with minimal amount of work and resources, focusing on material that is not protected by copyright and sources which could provide us with large chunks of text for each cleared permission. The two main sources considered were therefore official texts and texts from news media. Only digitally available texts are included in the corpus and formats that can be problematic are not processed. The corpus texts are morphosyntactically tagged and provided with metadata. Processes have been set up for continuous text collection, cleaning and annotation. The corpus is available for search and download with permissive licenses. The dataset is intended to be clearly versioned with the first version released in early 2018. Texts will be collected continually and a new version published every year.

Session P61 - Lexicon (3)

11th May 2018, 14:45

Chair person: **John McCrae** Poster Session

TriMED: A Multilingual Terminological Database

Federica Vezzani, Giorgio Maria Di Nunzio and Geneviève Henrot

Three precise categories of people are confronted with the complexity of medical language: physicians, patients and scientific translators. The purpose of this work is to develop a methodology for the implementation of a terminological tool that contributes to solve problems related to the opacity that characterizes communication in the medical field among its various actors. The main goals are: i) satisfy the peer-to-peer communication, ii) facilitate the comprehension of medical information by patients, and iii) provide a regularly updated resource for scientific translators. We illustrate our methodology and its application through the description of a multilingual terminological-phraseological resource named TriMED. This terminological database will consist of records designed to create a terminological bridge between the various registers (specialist, semi-specialist, non-specialist) as well as across the languages considered. In this initial analysis, we restricted to the field of breast cancer, and the terms to be analyzed will be extracted from a corpus in English, accompanied by all relevant linguistic information and properties, and re-attached to their pragmatic equivalent in Italian and French.

Preparation and Usage of Xhosa Lexicographical Data for a Multilingual, Federated Environment

Sonja Bosch, Thomas Eckart, Bettina Klimek, Dirk Goldhahn and Uwe Quasthoff

The South African linguistic landscape is characterised by multilingualism and the influence between their eleven official and some local languages. Unfortunately, for most of the languages the amount and quality of available lexicographical data is suboptimal, even though its availability is essential for all educational institutions and for the development of state-of-the-art language technology. In this paper we present a new source of lexicographical data for Xhosa, a language spoken by more than eight million speakers. For its utilisation in a multilingual and federated environment it is modelled using a dedicated OWL ontology for Bantu languages and possesses all features that are currently considered integral for the promotion of resource reuse as well as long-term usage. In the future, the introduced ontology may be used for other Bantu languages as well and may ease their combination to achieve more extensive, multilingual data stocks.

A Lexicon of Discourse Markers for Portuguese – LDM-PT

Amália Mendes, Iria Del Río Gayo, Manfred Stede and Felix Dombek

We present LDM-PT, a lexicon of discourse markers for European Portuguese, composed of 252 pairs of discourse marker/rhetorical sense. The lexicon covers conjunctions, prepositions, adverbs, adverbial phrases and alternative lexicalizations with a connective function, as in the PDTB (Prasad et al., 2008; Prasad et al., 2010). For each discourse marker in the lexicon, there is information regarding its type, category, mood and tense restrictions over the sentence it introduces, rhetorical sense, following the PDTB 3.0 sense hierarchy (Webber et al., 2016), as well as a link to an English near-synonym and a corpus example. The lexicon is compiled in a single excel spread sheet that is later converted to an XML scheme compatible with the DiMLex format (Stede, 2002). We give a detailed description of the contents and format of the lexicon, and discuss possible applications of this resource for discourse studies and discourse processing tools for Portuguese.

One Language to rule them all: modelling Morphological Patterns in a Large Scale Italian Lexicon with SWRL

Fahad Khan, Andrea Bellandi, Francesca Frontini and Monica Monachini

We present an application of Semantic Web Technologies to computational lexicography. More precisely we describe the publication of the \textit{morphological layer} of the Italian Parole Simple Clips lexicon (PSC-M) as linked open data. The novelty of our work is in the use of the Semantic Web Rule Language (SWRL) to encode morphological patterns, thereby allowing the automatic derivation of the inflectional variants of the entries in the lexicon. By doing so we make these patterns available in a form that is human readable and that therefore gives a comprehensive morphological description of a large number of Italian words.

Metaphor Suggestions based on a Semantic Metaphor Repository

Gerard De Melo

Metaphors are not only remarkably pervasive in both informal and formal text, but reflect fundamental properties of human cognition. This paper presents an algorithmic model that suggests metaphoric means of referring to concepts. For example, given a word such as "government", the method may propose expressions such as "father", "nanny", corresponding to common ways of thinking about the government. To achieve this, the model draws on MetaNet, a manually created repository of conceptual metaphor, in conjunction with lexical resources like FrameNet

and WordNet and automated interlinking techniques. These resources are connected and their overall graph structure allows us to propose potential metaphoric means of referring to the given input words, possibly constrained with additional metaphoric seed words, which may be provided as supplementary inputs. The experiments show that this algorithm greatly expands the potential of the repository.

The Linguistic Category Model in Polish (LCM-PL)

Aleksander Wawer and Justyna Sarzyńska

This article describes the first public release of Linguistic Category Model (LCM) dictionary for the Polish language (LCM-PL). It is used for verb categorization in terms of their abstractness and applied in many research scenarios, mostly in psychology. The dictionary consists of three distinctive parts: (1) sense-level manual annotation, (2) lexeme-level manual annotation, (3) lexeme-level automated annotation. The part (1) is of high quality yet the most expensive to obtain, therefore we complement it with options (2) and (3) to generate LCM labels for all verbs in Polish. Our dictionary is freely available for use and integrated with Słowosieć 3.0 (the Polish WordNet). Its quality will improve: we'll add more manually annotated senses and increase the quality of automated annotations.

WordNet-Shp: Towards the Building of a Lexical Database for a Peruvian Minority Language

Diego Maguiño Valencia, Arturo Oncevay and Marco Antonio Sobrevilla Cabezudo

WordNet-like resources are lexical databases with highly relevance information and data which could be exploited in more complex computational linguistics research and applications. The building process requires manual and automatic tasks, that could be more arduous if the language is a minority one with fewer digital resources. This study focuses in the construction of an initial WordNet database for a low-resourced and indigenous language in Peru: Shipibo-Konibo (shp). First, the stages of development from a scarce scenario (a bilingual dictionary shp-es) are described. Then, it is proposed a synset alignment method by comparing the definition glosses in the dictionary (written in Spanish) with the content of a Spanish WordNet. In this sense, word2vec similarity was the chosen metric for the proximity measure. Finally, an evaluation process is performed for the synsets, using a manually annotated Gold Standard in Shipibo-Konibo. The obtained results are promising, and this resource is expected to serve well in further applications, such as word sense disambiguation and even machine translation in the shp-es language pair.

Retrieving Information from the French Lexical Network in RDF/OWL Format

Alexsandro Fonseca, Fatiha Sadat and François Lareau

In this paper, we present a Java API to retrieve the lexical information from the French Lexical Network, a lexical resource based on the Meaning-Text Theory's lexical functions, which was previously transformed to an RDF/OWL format. We present four API functions: one that returns all the lexical relations between two given vocables; one that returns all the lexical relations and the lexical functions modeling those relations for two given vocables; one that returns all the lexical relations encoded in the lexical network modeled by a specific lexical function; and one that returns the semantic perspectives for a specific lexical function. This API was used in the identification of collocations in a French corpus of 1.8 million sentences and in the semantic classification of these collocations.

Session P62 - Named Entity Recognition

11th May 2018, 14:45

Chair person: **Gilles Francopoulo** Poster Session

Transforming Wikipedia into a Large-Scale Fine-Grained Entity Type Corpus

Abbas Ghaddar and Phillippe Langlais

This paper presents WiFiNE, an English corpus annotated with fine-grained entity types. We propose simple but effective heuristics we applied to English Wikipedia to build a large, high quality, annotated corpus. We evaluate the impact of our corpus on the fine-grained entity typing system of Shimaoka et al. (2017), with 2 manually annotated benchmarks, FIGER (GOLD) and ONTONOTES . We report state-of-the-art performances, with a gain of 0.8 micro F1 score on the former dataset and a gain of 2.7 macro F1 score on the latter one, despite the fact that we employ the same quantity of training data used in previous works. We make our corpus available as a resource for future works.

Error Analysis of Uyghur Name Tagging: Language-specific Techniques and Remaining Challenges

Halidanmu Abudukelimu, Adudoukelimu Abulizi, Boliang Zhang, Xiaoman Pan, Di Lu, Heng Ji and Yang Liu

Regardless of numerous efforts at name tagging for Uyghur, there is limited understanding on the performance ceiling. In this paper, we take a close look at the successful cases and perform careful analysis on the remaining errors of a state-of-the-art Uyghur name tagger, systematically categorize challenges, and propose possible solutions. We conclude that simply adopting a machine learning model which is proven successful for high-resource languages

along with language-independent superficial features is unlikely to be effective for Uyghur, or low-resource languages in general. Further advancement requires exploiting rich language-specific knowledge and non-traditional linguistic resources, and novel methods to encode them into machine learning frameworks.

BiLSTM-CRF for Persian Named-Entity Recognition ArmanPersoNERCorpus: the First Entity-Annotated Persian Dataset

Hanieh Poostchi, Ehsan Zare Borzeshi and Massimo Piccardi

Named-entity recognition (NER) can still be regarded as work in progress for a number of Asian languages due to the scarcity of annotated corpora. For this reason, with this paper we publicly release an entity-annotated Persian dataset and we present a performing approach for Persian NER based on a deep learning architecture. In addition to the entity-annotated dataset, we release a number of word embeddings (including GloVe, skip-gram, CBOW and Hellinger PCA) trained on a sizable collation of Persian text. The combination of the deep learning architecture (a BiLSTM-CRF) and the pre-trained word embeddings has allowed us to achieve a 77.45% CoNLL F1 score, a result that is more than 12 percentage points higher than the best previous result and interesting in absolute terms.

Data Anonymization for Requirements Quality Analysis: a Reproducible Automatic Error Detection Task

Juyeon Kang and Jungyeul Park

In this work, we aim at identifying potential problems of ambiguity, completeness, conformity, singularity and readability in system and software requirements specifications. Those problems arise particularly when they are written in a natural language. While we describe them from a linguistic point of view, the business impacts of each potential error are also considered in system engineering context. We investigate and explore error patterns for requirements quality analysis by manually analyzing the corpus. This analysis is based on the requirements grammar that we developed in our previous work. In addition, this paper extends our previous work in a two-fold way: (1) we increase more than twice the number of evaluation data (1K sentences) through a manual verification process, and (2) we anonymize all sensible and confidential entities in evaluation data to make our data publicly available. We also provide the baseline system using conditional random fields for requirements quality analysis, and we obtain 79.47\% for the F$_1$ score on proposed evaluation data.

A German Corpus for Fine-Grained Named Entity Recognition and Relation Extraction of Traffic and Industry Events

Martin Schiersch, Veselina Mironova, Maximilian Schmitt, Philippe Thomas, Aleksandra Gabryszak and Leonhard Hennig

Monitoring mobility- and industry-relevant events is important in areas such as personal travel planning and supply chain management, but extracting events pertaining to specific companies, transit routes and locations from heterogeneous, high-volume text streams remains a significant challenge. This work describes a corpus of German-language documents which has been annotated with fine-grained geo-entities, such as streets, stops and routes, as well as standard named entity types. It has also been annotated with a set of 15 traffic- and industry-related n-ary relations and events, such as accidents, traffic jams, acquisitions, and strikes. The corpus consists of newswire texts, Twitter messages, and traffic reports from radio stations, police and railway companies. It allows for training and evaluating both named entity recognition algorithms that aim for fine-grained typing of geo-entities, as well as n-ary relation extraction systems.

A Corpus Study and Annotation Schema for Named Entity Recognition and Relation Extraction of Business Products

Saskia Schön, Veselina Mironova, Aleksandra Gabryszak and Leonhard Hennig

Recognizing non-standard entity types and relations, such as B2B products, product classes and their producers, in news and forum texts is important in application areas such as supply chain monitoring and market research. However, there is a decided lack of annotated corpora and annotation guidelines in this domain. In this work, we present a corpus study, an annotation schema and associated guidelines, for the annotation of product entity and company-product relation mentions. We find that although product mentions are often realized as noun phrases, defining their exact extent is difficult due to high boundary ambiguity and the broad syntactic and semantic variety of their surface realizations. We also describe our ongoing annotation effort, and present a preliminary corpus of English web and social media documents annotated according to the proposed guidelines.

Portuguese Named Entity Recognition using Conditional Random Fields and Local Grammars

Juliana Pirovani and Elias Oliveira

Named Entity Recognition involves automatically identifying and classifying entities such as persons, places, and organizations, and it is a very important task in Information Extraction. Conditional Random Fields is a probabilistic method for structured prediction,

which can be used in this task. This paper presents the use of Conditional Random Fields for Named Entity Recognition in Portuguese texts considering the term classification obtained by a Local Grammar as an additional informed feature. Local grammars are handmade rules to identify named entities within the text. The Golden Collection of the First and Second HAREM considered as a reference for Named Entity Recognition systems in Portuguese were used as training and test sets respectively. The results obtained outperform the results competitive systems reported in the literature.

M-CNER: A Corpus for Chinese Named Entity Recognition in Multi-Domains

Qi Lu, YaoSheng Yang, Zhenghua Li, Wenliang Chen and Min Zhang

In this paper, we present a new corpus for Chinese Named Entity Recognition (NER) from three domains : human-computer interaction, social media, and e-commerce. The annotation procedure is conducted in two rounds. In the first round, one sentence is annotated by more than one persons independently. In the second round, the experts discuss the sentences for which the annotators do not make agreements. Finally, we obtain a corpus which have five data sets in three domains. We further evaluate three popular models on the newly created data sets. The experimental results show that the system based on Bi-LSTM-CRF performs the best among the comparison systems on all the data sets. The corpus can be used for further studies in research community.

SlugNERDS: A Named Entity Recognition Tool for Open Domain Dialogue Systems

Kevin Bowden, Jiaqi Wu, Shereen Oraby, Amita Misra and Marilyn Walker

In dialogue systems, the tasks of named entity recognition (NER) and named entity linking (NEL) are vital preprocessing steps for understanding user intent, especially in open domain interaction where we cannot rely on domain-specific inference. UCSC's effort as one of the funded teams in the 2017 Amazon Alexa Prize Contest has yielded Slugbot, an open domain social bot, aimed at casual conversation. We discovered several challenges specifically associated with both NER and NEL when building Slugbot, such as that the NE labels are too coarse-grained or the entity types are not linked to a useful ontology. Moreover, we have discovered that traditional approaches do not perform well in our context: even systems designed to operate on tweets or other social media data do not work well in dialogue systems. In this paper, we introduce Slugbot's Named Entity Recognition for dialogue Systems (SlugNERDS), a NER and NEL tool which is optimized to address these issues. We describe two new resources that we are building as part of this work: SlugEntityDB and SchemaActuator. We believe these resources will be useful for the research community.

Transfer Learning for Named-Entity Recognition with Neural Networks

Ji Young Lee, Franck Dernoncourt and Peter Szolovits

Recent approaches based on artificial neural networks (ANNs) have shown promising results for named-entity recognition (NER). In order to achieve high performances, ANNs need to be trained on a large labeled dataset. However, labels might be difficult to obtain for the dataset on which the user wants to perform NER: label scarcity is particularly pronounced for patient note de-identification, which is an instance of NER. In this work, we analyze to what extent transfer learning may address this issue. In particular, we demonstrate that transferring an ANN model trained on a large labeled dataset to another dataset with a limited number of labels improves upon the state-of-the-art results on two different datasets for patient note de-identification.

Session P63 - Parsing, Syntax, Treebank (2)

11th May 2018, 14:45

Chair person: **Jan Hajič** Poster Session

ForFun 1.0: Prague Database of Forms and Functions – An Invaluable Resource for Linguistic Research

Marie Mikulová and Eduard Bejček

In this paper, we introduce the first version of ForFun, Prague Database of Forms and Functions, as an invaluable resource for profound linguistic research, particularly in describing syntactic functions and their formal realizations. ForFun is built with the use of already existing richly syntactically annotated corpora, collectively called Prague Dependency Treebanks. ForFun brings this complex annotation of Czech sentences closer to researchers. We demonstrate that ForFun 1.0 provides valuable and rich material allowing to elaborate various syntactic issues in depth. We believe that nowadays when corpus linguistics differs from traditional linguistics in its insistence on a systematic study of authentic examples of language in use, our database will contribute to the comprehensive syntactic description.

The LIA Treebank of Spoken Norwegian Dialects

Lilja Øvrelid, Andre Kåsen, Kristin Hagen, Anders Nøklestad, Per Erik Solberg and Janne Bondi Johannessen

This article presents the LIA treebank of transcribed spoken Norwegian dialects. It consists of dialect recordings made in the period between 1950–1990, which have been digitised,

transcribed, and subsequently annotated with morphological and dependency-style syntactic analysis as part of the LIA (Language Infrastructure made Accessible) project at the University of Oslo. In this article, we describe the LIA material of dialect recordings and its transcription, transliteration and further morphosyntactic annotation. We focus in particular on the extension of the native NDT annotation scheme to spoken language phenomena, such as pauses and various types of disfluencies, and present the subsequent conversion of the treebank to the Universal Dependencies scheme. The treebank currently consists of 13,608 tokens, distributed over 1396 segments taken from three different dialects of spoken Norwegian. The LIA treebank annotation is an on-going effort and future releases will extend on the current data set.

Errator: a Tool to Help Detect Annotation Errors in the Universal Dependencies Project

Guillaume Wisniewski

Enforcing guidelines compliance is today one of the main challenge faced by the Universal Dependencies project. This work introduces ERRATOR, a set of tools implementing the annotation variation principle that can be used to help annotators find and correct errors in the different layers of annotations of UD treebanks. The results of a first annotation campaign that used ERRATOR to correct and harmonize the annotations of the different French corpora are also described.

SandhiKosh: A Benchmark Corpus for Evaluating Sanskrit Sandhi Tools

Shubham Bhardwaj, Neelamadhav Gantayat, Nikhil Chaturvedi, Rahul Garg and Sumeet Agarwal

Sanskrit is an ancient Indian language. Several important texts which are of interest to people all over the world today were written in Sanskrit. The Sanskrit grammar has a precise and complete specification given in the text Astadhyayi by Panini. This has led to the development of a number of {\em Sanskrit Computational Linguistics} tools for processing and analyzing Sanskrit texts. Unfortunately, there has been no effort to standardize and critically validate these tools. In this paper, we develop a Sanskrit benchmark called SandhiKosh to evaluate the completeness and accuracy of Sanskrit Sandhi tools. We present the results of this benchmark on three most prominent Sanskrit tools and demonstrate that these tools have substantial scope for improvement. This benchmark will be freely available to researchers worldwide and we hope it will help everyone working in this area evaluate and validate their tools.

Czech Legal Text Treebank 2.0

Vincent Kríž and Barbora Hladka

The Czech Legal Text Treebank 2.0 (CLTT 2.0) contains texts that come from the legal domain and are manually syntactically annotated. The syntactic annotation in CLTT 2.0 is more elaborate than in CLTT 1.0. In addition, CLTT 2.0 contains two new annotation layers, namely the layer of entities and the layer of semantic entity relations. In total, CLTT 2.0 consists of two documents, 1,121 sentences and 40,950 tokens.

Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU

Normunds Gruzitis, Lauma Pretkalnina, Baiba Saulite, Laura Rituma, Gunta Nespore-Berzkalne, Arturs Znotins and Peteris Paikens

This paper presents a work in progress to create a multilayered syntactically and semantically annotated text corpus for Latvian. The broad application area we address is natural language understanding (NLU), while more specific applications are abstractive text summarization and knowledge base population, which are required by the project industrial partner, Latvian information agency LETA, for the automation of various media monitoring processes. Both the multilayered corpus and the downstream applications are anchored in cross-lingual state-of-the-art representations: Universal Dependencies (UD), FrameNet, PropBank and Abstract Meaning Representation (AMR). In this paper, we particularly focus on the consecutive annotation of the treebank and framebank layers. We also draw links to the ultimate AMR layer and the auxiliary named entity and coreference annotation layers. Since we are aiming at a medium-sized still general-purpose corpus for a less-resourced language, an important aspect we consider is the variety and balance of the corpus in terms of genres, authors and lexical units.

Test Sets for Chinese Nonlocal Dependency Parsing

Manjuan Duan and William Schuler

Chinese is a language rich in nonlocal dependencies. Correctly resolving these dependencies is crucial in understanding the predicate-argument structure of a sentence. Making full use of the trace annotations in the Penn Chinese Treebank, this research contributes several test sets of Chinese nonlocal dependencies which occur in different grammatical constructions. These datasets can be used by an automatic dependency parser to evaluate its performance on nonlocal dependency resolution in various syntactic constructions in Chinese.

Adding Syntactic Annotations to Flickr30k Entities Corpus for Multimodal Ambiguous Prepositional-Phrase Attachment Resolution

Sebastien Delecraz, Alexis Nasr, FREDERIC BECHET and Benoit Favre

We propose in this paper to add to the captions of the Flickr30k Entities corpus some syntactic annotations in order to study the joint processing of image and language features for the Preposition-Phrase attachment disambiguation task. The annotation has been performed on the English version of the captions and automatically projected on their French and German translations.

Analyzing Middle High German Syntax with RDF and SPARQL

Christian Chiarcos, Benjamin Kosmehl, Christian Fäth and Maria Sukhareva

The paper presents technological foundations for an empirical study of Middle High German (MHG) syntax. We aim to analyze the diachronic changes of MHG syntax on the example of direct and indirect object alterations in the middle field. In the absence of syntactically annotated corpora, we provide a rule-based shallow parser and an enrichment pipeline with the purpose of quantitative evaluation of a qualitative hypothesis. % It discusses how quantitative evaluation of qualitative methods can be used to open up a new prospective on a well-studied phenomenon. We provide a publicaly available enrichment and annotation pipeline grounded. A technologically innovative aspect is the application of CoNLL-RDF and SPARQL Update for parsing.

Cheating a Parser to Death: Data-driven Cross-Treebank Annotation Transfer

Djamé Seddah, Eric De La Clergerie, Benoît Sagot, Héctor Martínez Alonso and Marie Candito

We present an efficient and accurate method for transferring annotations between two different treebanks of the same language. This method led to the creation of a new instance of the French Treebank (Abeillé et al., 2003), which follows the Universal Dependency annotation scheme and which was proposed to the participants of the CoNLL 2017 Universal Dependency parsing shared task (Zeman et al., 2017). Strong results from an evaluation on our gold standard (94.75% of LAS, 99.40% UAS on the test set) demonstrate the quality of this new annotated data set and validate our approach.

Universal Dependencies and Quantitative Typological Trends. A Case Study on Word Order

Chiara Alzetta, Felice Dell'Orletta, Simonetta Montemagni and Giulia Venturi

The paper presents a new methodology aimed at acquiring typological evidence from "gold" treebanks for different languages. In particular, it investigates whether and to what extent algorithms developed for assessing the plausibility of automatically produced syntactic annotations could contribute to shed light on key issues of the linguistic typological literature. It reports the first and promising results of a case study focusing on word order patterns carried out on three different languages (English, Italian and Spanish).

Session P64 - Wordnets and Ontologies

11th May 2018, 14:45

Chair person: **Elena Montiel-Ponsoda** Poster Session

Undersampling Improves Hypernymy Prototypicality Learning

Koki Washio and Tsuneaki Kato

This paper focuses on supervised hypernymy detection using distributional representations for unknown word pairs. Levy et al. (2015) demonstrated that supervised hypernymy detection suffers from overfitting hypernyms in training data. We show that the problem of overfitting on this task is caused by a characteristic of datasets, which stems from the inherent structure of the language resources used, hierarchical thesauri. The simple data preprocessing method proposed in this paper alleviates this problem. To be more precise, we demonstrate through experiments that the problem that hypernymy classifiers overfit hypernyms in training data comes from a skewed word frequency distribution brought by the quasi-tree structure of a thesaurus, which is a major resource of lexical semantic relation data, and propose a simple undersampling method based on word frequencies that can effectively alleviate overfitting and improve distributional prototypicality learning for unknown word pairs.

Interoperability of Language-related Information: Mapping the BLL Thesaurus to Lexvo and Glottolog

Vanya Dimitrova, Christian Fäth, Christian Chiarcos, Heike Renner-Westermann and Frank Abromeit

Since 2013, the thesaurus of the Bibliography of Linguistic Literature (BLL Thesaurus) has been applied in the context of the Lin|gulis|tik portal, a hub for linguistically relevant information. Several consecutive projects focus on the modeling of the

BLL Thesaurus as ontology and its linking to terminological repositories in the Linguistic Linked Open Data (LLOD) cloud. Those mappings facilitate the connection between the Lin|gu|is|tik portal and the cloud. In the paper, we describe the current efforts to establish interoperability between the language-related index terms and repositories providing language identifiers for the web of Linked Data. After an introduction of Lexvo and Glottolog, we outline the scope, the structure, and the peculiarities of the BLL Thesaurus. We discuss the challenges for the design of scientifically plausible language classification and the linking between divergent classifications. We describe the prototype of the linking model and propose pragmatic solutions for structural or conceptual conflicts. Additionally, we depict the benefits from the envisaged interoperability - for the Lin|gu|is|tik portal, and the Linked Open Data Community in general.

Browsing and Supporting Pluricentric Global Wordnet, or just your Wordnet of Interest

António Branco, Ruben Branco, Chakaveh Saedi and João Silva

In this paper we proceed with a systematic gathering of design requirements for wordnet browsers that permit to consult the content of wordnets. This is undertaken together with a review of the functionalities of existing browsers. On the basis of this analysis, we present a new wordnet browser we developed that meets these requirements and thus complies with the most ample range of design features. This is an open source browser that is freely distributed and can be reused by anyone interested in doing research on or just using wordnets. We also introduce the notion of a pluricentric global wordnet, for whose undertaking this new advanced browser appears as an important instrument and motivation. This is a promising operative conception for a bootstrapped yet effective process towards the ultimate global wordnet, where all individual wordnets from all languages are meant to eventually converge together, in spite of the plurality of their formats, licenses, depth, etc. that is intrinsic to an inherently plural endeavor undertaken by multiple actors under multiple constraints across the world.

Cross-checking WordNet and SUMO Using Meronymy

Javier Alvez, Itziar Gonzalez-Dios and German Rigau

We report on the practical application of a black-box testing methodology for the validation of the knowledge encoded in WordNet, SUMO and their mapping by using automated theorem provers. Our proposal is based on the part-whole information provided by WordNet, out of which we automatically create a large set of tests. Our experimental results confirm that the proposed system enables the validation of some pieces of information and also the detection of missing information or inconsistencies among these resources.

Extended HowNet 2.0 – An Entity-Relation Common-Sense Representation Model

Wei-Yun Ma and Yueh-Yin Shih

In this paper, we propose Extended HowNet 2.0 – an entity-relation common-sense representation model. Comparing to HowNet and Extended HowNet, E-HowNet 2.0 has the following improvements: (a) Reorganizing the hierarchical structure of primitives and basic concepts; (b) Rich lexical information: In addition to sense definition, each entry of lexical sense may also include operational expressions as well as semantic functions which facilitate future semantic composition processes. (c) Improvement of sense definitions and sense definitions for basic concepts. (d) Developing a new automatic ontology reconstruction system. (e) Developing a query system called E-HowNet Relation Database for flexibly clustering concepts.We hope Extended HowNet 2.0 can bring significant benefits to the community of lexical semantics and natural language understanding.

The Circumstantial Event Ontology (CEO) and ECB+/CEO: an Ontology and Corpus for Implicit Causal Relations between Events

Roxane Segers, Tommaso Caselli and Piek Vossen

In this paper, we describe the Circumstantial Event Ontology (CEO), a newly developed ontology for calamity events that models semantic circumstantial relations between event classes, where we define circumstantial as inferred implicit causal relations. The circumstantial relations are inferred from the assertions of the event classes that involve a change to the same property of a participant. Our model captures that the change yielded by one event, explains to people the happening of the next event when observed. We describe the meta model and the contents of the ontology, the creation of a manually annotated corpus for circumstantial relations based on ECB+ and the first results on the evaluation of the ontology.

Profiling Medical Journal Articles Using a Gene Ontology Semantic Tagger

Mahmoud El-Haj, Paul Rayson, Scott Piao and Jo Knight

In many areas of academic publishing, there is an explosion of literature, and sub-division of fields into subfields, leading to stove-piping where sub-communities of expertise become disconnected from each other. This is especially true in the genetics literature over the last 10 years where researchers are no longer able to maintain knowledge of previously related areas. This paper extends several approaches based on natural language

processing and corpus linguistics which allow us to examine corpora derived from bodies of genetics literature and will help to make comparisons and improve retrieval methods using domain knowledge via an existing gene ontology. We derived two open access medical journal corpora from PubMed related to psychiatric genetics and immune disorder genetics. We created a novel Gene Ontology Semantic Tagger (GOST) and lexicon to annotate the corpora and are then able to compare subsets of literature to understand the relative distributions of genetic terminology, thereby enabling researchers to make improved connections between them.

Towards a Conversation-Analytic Taxonomy of Speech Overlap

Felix Gervits and Matthias Scheutz

We present a taxonomy for classifying speech overlap in natural language dialogue. The scheme classifies overlap on the basis of several features, including onset point, local dialogue history, and management behavior. We describe the various dimensions of this scheme and show how it was applied to a corpus of remote, collaborative dialogue. Moving forward, this will serve as the basis for a computational model of speech overlap, and for use in artificial agents that interact with humans in social settings.

Indian Language Wordnets and their Linkages with Princeton WordNet

Diptesh Kanojia, Kevin Patel and Pushpak Bhattacharyya

Wordnets are rich lexico-semantic resources. Linked wordnets are extensions of wordnets, which link similar concepts in wordnets of different languages. Such resources are extremely useful in many Natural Language Processing (NLP) applications, primarily those based on knowledge-based approaches. In such approaches, these resources are considered as gold standard/oracle. Thus, it is crucial that these resources hold correct information. Thereby, they are created by human experts. However, human experts in multiple languages are hard to come by. Thus, the community would benefit from sharing of such manually created resources. In this paper, we release mappings of 18 Indian language wordnets linked with Princeton WordNet. We believe that availability of such resources will have a direct impact on the progress in NLP for these languages.

Authors Index