EIGHTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION

Held under the Patronage of Ms Neelie Kroes, Vice-President of the European Commission, Digital Agenda Commissioner

MAY 23-24-25, 2012

ISTANBUL LÜTFI KIRDAR CONVENTION & EXHIBITION CENTRE
ISTANBUL, TURKEY

CONFERENCE ABSTRACTS

Editors: Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis.

Assistant Editors: Hélène Mazo, Sara Goggi, Olivier Hamon

LREC 2012, EIGHTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION

Title: LREC 2012 Conference Abstracts

Distributed by:
ELRA – European Language Resources Association
55-57, rue Brillat Savarin
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 978-2-9517408-7-7
EAN 9782951740877

Introduction of the Conference Chair
Nicoletta Calzolari

I wish first to express to Ms Neelie Kroes, Vice-President of the European Commission, Digital agenda Commissioner, the gratitude of the Program Committee and of all LREC participants for her Distinguished Patronage of LREC 2012.

Even if every time I feel we have reached the top, this 8th LREC is continuing the tradition of breaking previous records: this edition we received 1013 submissions and have accepted 697 papers, after reviewing by the impressive number of 715 colleagues. We have accepted 30 Workshops and 10 Tutorials. We can't deny that the field of Language Resources and Evaluation is flourishing more than ever! So far more than 1100 people have already registered. From all these signals we see that LREC continues to be – as many say – "the conference where you have to be and where you meet everyone".

The high acceptance rate is an important characteristic that is at the essence of LREC. We consider of utmost importance to provide a global view of the current trends, in all their dimensions and as reflected in many languages. Multilingualism is a core feature of the field and it is important to show how the field is evolving not only with respect to new methodologies and algorithms but also with respect to more advanced treatment of more and more languages. This is also a strategic choice underlying the importance of safeguarding the heritage of world's linguistic diversity.

In the preparation of the program, while trying to arrange all the pieces of the big puzzle, it is a pleasure touching the hot topics of these days. It is always extremely interesting to spot the major changes with respect to previous editions and monitor the evolution of the field.

Major trends, i.e. not the most crowded topics but those increasing with respect to last LREC:

- "Data", as normal in our conference, but I would say even more than before: data/corpora (in all the modalities and for many purposes: annotation, extraction, classification, translation, and so on) receive even more attention than last time.
- Machine Translation and multilingualism is a major topic, on which a lot of work is carried out.
- Infrastructural initiatives, strategies, national and international projects are a big issue, as usual inside the LREC community.
- Tools and systems for text analysis at many levels are presented in many papers.
- Temporal and spatial information is also relatively increasing.

New entries, i.e. topics emerging this year:

- New media (twitter, chats and the like)
- Crowdsourcing
- Child language corpora

More or less stable, at the same level as last year are topics such as:

- Semantics and Knowledge, in all their variations: from annotation of anaphoric information, to ontologies and lexicons, disambiguation, named entities recognition, information extraction, to mention just a few.
- Subjectivity, declined in various nuances: emotions, opinions, and sentiments.
- Dialogue and discourse, with contributions from both the Speech and Text communities.
- Speech and Multimodal resources, tools, systems.
- And finally, evaluation and validation methodologies, as an important part of quite many papers.

A slightly declining tendency seems to be associated with:

- Lexical resources
- Grammar, syntax, parsing

Are these solved issues?

As usual, a distinctive feature of LREC is the emphasis given to infrastructural and strategic initiatives. I consider this a very important characteristic of the Language Resource field, one that ELRA has always supported, and one that deserves some reflection. This is probably due to the recognised fact of the necessity to work on massive amounts of data, moreover multiplied by the many languages, and to the infrastructural nature of language resources. The fact that ours is a data-intensive discipline requires building on each other results if we want to achieve something serious, and requires joining efforts. This recognition has led in the last years to the creation of important infrastructures, such as in Europe META-SHARE and CLARIN-ERIC.

We must now seriously think at how to enter with force into the area of "big data" – and also "open data" – the next frontier for competition and innovation, where we have to cope with very strong and organised communities. But we must also learn from them, in particular how to enable working together in huge experiments with thousands of colleagues cooperating on common objectives. This is, I believe, our next step if we want to address the challenges of big data and achieve the status of a mature science. And this requires networking, collaboration and sharing, of ideas and data. I hope LREC helps going in this direction.

LREC Innovations

We continue also the tradition of introducing some innovation at LREC.

- The ***LRE Map***, which started only two years ago, is already established, consulted every day and is used in other major conferences. At this LREC we collected descriptions for more than 1200 resources in more than 200 languages! More details on the Map will be found at the ELRA booth where it is presented.
- The novelty of this year is the ***Language Library***. It is an experiment to see if we can set up a platform for collaborative work on the processing (annotation, translation, ...) of language resources. The Library will be presented in the ELRA booth as well. Both the LRE Map and the Library have been conceived as services for the community and are being built by the community itself.
- Also this time we have introduced in the program two "***Special sessions***" on emerging new topics – New Media and Crowdsourcing – with a slot dedicated to general discussion.
- We repeat this year the experience of the ***EC Village***, so successful last time. We have even more booths, representing a very large number of EC projects.
- A new insertion of this LREC is the ***EC Track*** on the second day of the conference, organised by Roberto Cencioni, where some EC projects – and related trends – in the field of Language Resources and Evaluation are presented and discussed. This, together with the EC Village, offers the possibility of getting a comprehensive picture of the EC initiatives in the field.

Acknowledgments

And finally, I wish to express my appreciation to all those who made this LREC possible and hopefully successful.

I first thank the Program Committee members, not only for their dedication in the huge task of selecting the papers, but also for the constant involvement in the various aspects around LREC. A particular thanks goes to Jan Odijk, who has been so helpful in the preparation of the program. And obviously to Khalid Choukri, who is in charge of so many organisational aspects around LREC.

I thank ELRA, which is the promoter of LREC, its own conference.

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

A particular thanks goes to the Local Committee, and especially to Mehmed Özkan (its chair): they have worked hard for many months to find the best solutions to local issues.

I express my gratitude to the Sponsors that have believed in the importance of our conference, and have helped with financial support. I am grateful to the authorities, and all associations, organisations, companies that have supported LREC in various ways, for their important cooperation.

I thank the workshop and tutorial organisers, who complement LREC of so many interesting events. A big thanks goes to all the authors, who provide the "substance" to LREC, and give us such a broad picture of the field.

I finally thank the two institutions that have dedicated such a great effort to this LREC, as to the previous ones, i.e. ELDA in Paris and my group at ILC-CNR in Pisa. Without their commitment LREC would not have been possible. The last, but not least, thanks are thus for: Hélène Mazo and Sara Goggi, two pillars of LREC without whose commitment for many months LREC would not happen, and the others who have helped and will help during the conference: Victoria Arranz, Cécile Barbier, Paola Baroni, Roberto Bartolini, Riccardo Del Gratta, Francesca Frontini, Olivier Hamon, Valérie Mapelli, Vincenzo Parrinelli, Valeria Quochi, Caroline Rannaud, Irene Russo, Priscille Schneller. We have solved together the many big and small problems of such a large conference. They will also assist you during the conference.

Now LREC is yours. You – the participants – are the real protagonist of LREC, you will make this LREC great. So, at the very end, my greatest thanks go to you all. I may not be able to speak with each one of you during the Conference (although I'll try!). I hope that you learn something, that you perceive and touch the excitement, fervour and liveliness 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.

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

The tradition of holding LREC in wonderful locations with a Mediterranean flavour continues, and Istanbul is a perfect LREC location! I am sure you will enjoy Istanbul during the LREC week. And I hope that Istanbul will enjoy the invasion of LRECers!

With all the Programme Committee, I welcome you at LREC 2012 in such a wonderful country as Turkey and wish you a fruitful Conference.

Enjoy LREC 2012 in Istanbul!

Nicoletta Calzolari
Chair
8th International Conference on Language Resources and Evaluation

Message from ELRA President
Stelios Piperidis

Welcome to LREC2012! Welcome to Istanbul!

Let me first express, on behalf of the ELRA Board and Members, our profound gratitude to Mrs Neelie Kroes, Vice-President of the European Commission, for her Distinguished Patronage of LREC 2012 and for honoring us with her message.

The 8th edition of LREC takes place in a most interesting context for our field, in times when everything is changing at a dazzling speed. The common denominator of all changes being the quest of the best recipe for rapid development, competitiveness and innovation. Information technology at large has an undisputable contribution to this endeavour, as access to the whole spectrum from data, information, content, up to knowledge, and the ability to process it is one of the enabling factors leading to development and innovation. Taking into account the importance of language in this spectrum, language technologies have a decisive role to play. To be able to fulfill the role, understand the complex phenomena, tackle open problems, build useful applications and foster innovation, access to the necessary resources infrastructure is indispensable. It is, thus, not to wonder why during the last years issues relating to data access, open data, collaboration and exchange have absorbed a non-negligible portion of the energy of the scientific world and certainly of our field.

ELRA, established already in 1995, has played a pioneering role in this respect. At times when data-driven techniques had just started rising, and numerical and learning methods had not prevailed in the language technology field, ELRA was set up as an association that would take care of identifying, archiving, and distributing language data, as well as cater for language technology evaluation. These initial goals were soon extended into production and validation of language data, production of technology evaluation packages, as well as into offerings of specific services and organisation of specific events, LREC being the major such event since 1998.

ELRA has achieved to establish itself as a high-quality data centre, as a main player in the field of language resources and evaluation, enjoying a steady base of membership. Throughout its life, the Association has been constantly responding to new challenges and needs of the field, as well as adapting to new cultures and trends. Besides its ever evolving catalogue of language data with clear distribution rights and licences, ELRA has set up the Universal Catalogue, a bottom-up, community-built resource radar, a catalogue of resources, data and tools, irrespective of whether these can be acquired through ELDA, ELRA's Distribution Agency, or not. It supports the LRE Map, a concerted effort towards a new culture, a rich, community-built, information base, already embraced and re-used by many conferences, a unique tool for monitoring progress, identifying specific gaps, highlighting trends. It also supports the emerging Language Library, a platform for collaboratively built annotated resources at all levels, available as open resources to the whole community. A big "thank you" to all participants for contributing to this initiative.

In a similar vein, through its catalogues and other initiatives, ELRA reinforces its cooperation links with major data centres around the world, notably LDC and NICT, trying to unify all component catalogues into a Global Inventory of appropriately identified LRs.

To achieve its goals, ELRA has restructured its activities along three axes : a) *infrastructural*, taking care of identification, cataloguing, updating metadata-based LR description and documentation, and new simpler distribution mechanisms, b) *scientific*, taking care of LRs production & validation, and evaluation, and c) *promotion*, taking care of marketing and promotion activities like LREC and information aggregation services.

Along these axes, ELRA has currently set 5 main priorities:

- Opening up segments of its catalogue

Anticipating users' expectations, ELRA has decided to offer a large number of resources for free for the research community. Such an offer will consist of several sets of speech, text, multimodal databases that will be released for free regularly, as soon as legal aspects are cleared.

- Fostering the use of Public Sector Information

Being among the first to recognize the importance of public sector information, the Association joins forces with all stakeholders in the effort to reinforce and extend the free use of public sector information for research, technology and application development.

- Supporting META-SHARE, the new resource sharing and exchange infrastructure

ELRA, through its operational body ELDA, is a founding member of META-NET, and plays an important role in META-SHARE offering a range of services (among others supporting the legal helpdesk and the non-local repository of the infrastructure) and helping sustain it. Most of the resources already distributed by ELRA are also available on META-SHARE.

- Promoting collaborative and crowd-sourcing based methods of language resources building

During the last year, the Association has been investigating the setup of a dedicated platform for crowd-sourcing based language resource building. Your feedback through the questionnaire we have prepared is most valuable.

- Establishing the Language Resources and Evaluation Forum (LRE-F)

ELRA is establishing the Forum that is expected to help maintain and extend our vibrant community through sharing, exchange and collaboration assisting services.

Just a few words about the Antonio Zampolli Prize, the prize created by the ELRA Board in order to honour our founder and first president who did so much for the field of language resources. Citing the Prize articles: "The Antonio Zampolli Prize is intended to recognize the outstanding contributions to the advancement of Human Language Technologies through all issues related to Language Resources and Evaluation. In awarding the prize we are seeking to reward and encourage innovation and inventiveness in the development and use of language resources and evaluation of HLTs. The prize covers the field of Language Resources and Language Technology Evaluation in the areas of spoken language, written language and terminology". At the LREC 2012 conference, the Prize will be awarded for the fifth time. The ELRA Board has been very happy to receive very strong nominations made by outstanding people in the field, and we do recognize there are several persons who are eligible for this prestigious prize.

I would like to take the opportunity to thank all those who have worked so hard to make this conference a big success: the LREC Programme Committee, chaired by Nicoletta Calzolari, the Scientific Committee, the International Advisory Committee, the group in Pisa, Khalid Choukri and the ELDA staff in Paris, the two LREC "pillars" Helene Mazo and Sara Goggi, Mehmed Özkan and the Local Organising Committee. Each one of them in his/her own role has been taking care of the myriad of issues that pop up when undertaking the organisation of such a complex and demanding conference as LREC, a Herculean task. Particular thanks to all sponsors and supporters, all conference participants, workshop and tutorial organizers, project consortia participating in the EC Village; you have all helped outperform once again all previous editions in all the dimensions of our LREC conference.

Dear LREC Participants, The Conference is now in your hands. With your active participation in the oral sessions, your lively discussions with the presenters at the poster sessions, your visits to the EC Village and participation in the EC Track to discuss with the investigators of the research projects funded by the European Commission, LREC 2012 will be yet another success.

I welcome you all to LREC 2012 in magnificent Istanbul and wish you a fruitful conference.

Stelios Piperidis
ELRA President

Message from ELRA Secretary General and ELDA Managing Director
Khalid Choukri

Welcome to Istanbul and LREC 2012,

Welcome to this LREC 2012, the 8th edition of one of the major events in language sciences and technologies and the most visible service of ELRA to the community.

I would like to extend our warm welcome to the 140 representatives of ELRA members, attending LREC2012.

On behalf of ELRA members and LREC participants, I would like express our gratitude to Ms Neelie Kroes, Vice-President of the European Commission, in charge of the Digital agenda, for her Distinguished Patronage of LREC 2012.

Organizing LREC 2012 under the auspices of these distinguish patrons is an important sign, for us who manage signs, symbols and semantics, regarding the importance conferred to languages, multilingualism, information technologies and all related fields.

These issues are at the heart of EU Digital Agenda, an Agenda that should consider Language Technologies as an essential path to pave the way to automating not only human-machine interactions, human access to information but also human-human communications, across languages and across cultures.

After having organized LREC in Marrakech and Malta, two representatives of Semitic languages (Arabic, Maltese), we are this time in a city that played one of the most noticeable roles in forging Europe, parts of Asia and Africa history and geopolitics, as well as languages, with its own language family, the Turkic languages. After a number of centuries, during which Turkish shared many aspects with Arabic and Persian including a writing system, the foundation of the republic of Turkey came with the script reform (shifting from "Arabic" characters to Latin ones) and the foundation of the Turkish Language Association in 1932 under the patronage of Mustapha Kemal Ataturk himself. The association revived so many Turkic terms and came out with so many neologisms to establish the modern language. This experience, event if not unique in mankind history, is an important process for us, Language scientists and engineers.

At ELRA, we are very happy to carry out and support activities that help all languages to have access to resources essential for their move forward for a bright future and in particular for ensuring access to the digital world and reducing the digital divide.

We are very proud to organize this 8th LREC in that context and for that purpose: to offer our Community the forum it needs, where all players can meet and discuss hot issues related to language resources, technology evaluation, and language sciences.
With more than a thousand participants attending each LREC since 2008, we feel confident that such event where players from Academia and Industry can meet, where new comers, students and junior researchers can find background knowledge and where researchers can review new theories and trends.
With more than 1100 registered participants, more than 30 specialized workshops, about 10 tutorials, almost 700 papers at the main conference, we feel that the achievement is worth the effort we dedicate to make it happen. Boosted by this vitality and energy of our field, ELRA is moving forward with new

objectives and new services to anticipate the community expectations in its challenging task to bring in more supporting tools and automations, to overcome the language and cultural barriers, and help humans enjoy the multilingualism, multiculturalism of the global world of today and tomorrow.

Over the last 17th years (1995-2012), ELRA, driven by its members' instructions, requirements, expectations, has established a number of activities to serve them. LREC is "only" and (probably) the (most) visible aspect of such services.

As many of you know, the core activity of ELRA has been and continues to be identification of valuable Language Resources, useful for research, development and evaluation of Language Technologies. Such identification, followed by a time consuming process of negotiating distribution conditions and clearing all legal issues, led to the constitution of the ELRA catalogue of over 1000 language resources and evaluation packages

In order to help enrich such catalogue, ELRA initiated an identification process to collect and compile data on all existing resources, worldwide, to ensure that such information is shared within the community. This is our Universal Catalogue (UC). UC comprises all identified resources and a priority list is drawn before to launch the negotiations with right holders on sharing and distributing them.

To supplement this, another initiative was launched by ELRA at LREC'2010, the LRE'Map (a Language Resources and Evaluation map). LRE'Map allows each LREC author to describe resources used in his/her work. More than 1200 LR descriptions have been collected at this LREC. LRE'Map feature is now exploited by other conferences and we hope it will become a common feature to all Language Technology events (www.resourcesbook.eu). Such map contributes to spreading and sharing knowledge about LRs.

It is clear that such repositories and resources, along with fair, easy to use, and trustable legal conditions played a role in deployment of Languages Technology applications.

Since 2010, as partially reported on at LREC 2010, ELRA, through its operational body ELDA, is taking part to the META-NET Network of Excellence (Technologies for the Multilingual European Information Society). The main objective is to move forward and extend existing distribution and sharing mechanisms within a new paradigm. For this purpose, the consortium focuses on "Building an Open Resource Infrastructure", for sharing language resources and tools, referred to as META-SHARE.
META-SHARE aims to be "*a sustainable network of repositories of language data, tools and related web services documented with high-quality metadata, aggregated in central inventories allowing for uniform search and access to resources.*" (cf. http://www.meta-net.eu/meta-share).
One of the essential tasks of the project is related to the metadata issues with respect to the description of LRs. Work has being carried out for the specification of a metadata schema which builds upon available schemas e.g. ELRA, knowledge and expertise and provides a unified schema capable of handling the requirements of the community. These requirements comprise both the description of Language Resources and that of tools or technologies. A large number of Language Technology organizations have been debating the harmonization of such descriptions. In addition, this work aims to consider new modalities such as video and image (for e.g. sign languages, multi-sensor or multi-modal data, etc.). This work on metadata is now mature enough to be considered for standardization. More than 50 players have adopted it and many tools (metadata editor, converters from existing schemas, etc.) are made widely available.

A related issue on which ELRA and a large number of Language Technologies organization have been debating is the harmonization of the identification of LRs. A consensus seems to emerge regarding the set-up of a small executive committee, steered by a commission representing all key players in the field, data centers (ELRA, LDC, Allagin,/GSK, C-LDC,…), and the stack holders (ACL, IAMT, ISCA,…), to assign each LR an International Standard Language Resource Number (ISLRN), independently of whether the LR is accessible on Internet, Intranet, available or not, etc… whether it

has a DOI, a local PId, etc. Such ISLRN should guarantee that all LR usable within our field get a unique identifier that can be used to distinguish it from others.

Another important aspect, the harmonization of existing licensing schemas and the legal aspects, has been part of the discussions and in particular, ELDA focused on the commonalities between ELRA licenses and the ones promoted by Creative Commons, with the intention to harmonize such licenses under the new umbrella of META-SHARE, which was done and will be debated during this LREC at a dedicated workshop.

A version of the META-SHARE network of repositories is already available (www.meta-share.eu) and more information about it is provided in the ELRA's president message as well as at the corresponding LREC workshop, tutorial, and several accepted papers.

As indicated above, a major barrier that hinders the sharing of language resources and tools is the copyright and other IPR issues. ELRA and the META-SHARE partners have been working hard to offer a harmonized set of licenses that cover all needs and sharing/distributing scenarios. In parallel, ELRA continues to advocate for simplifying copyright and IPR issues concerning LR, in particular when used for research purposes. Such exception, which exists in a number of countries (e.g. section 107 of the US copyright law), deserves to be harmonized and extended to all countries. LREC offers a useful forum for debating such issue and hopefully coming up with a common declaration on this and other similar hot topics, to be pushed forward by all of us back home.

This has been a strong credo of the FLaReNet project (in which ELRA Board members and many stack holders including ELDA) took an active role. FLaReNet conclusions at its annual forums, advocated for this harmonization. It went beyond that and compiled a useful but critical roadmap, available to all (www.flarenet.eu), and drew a clear picture of the new trends and important expectations and paved the path for ELRA activities for the coming years. Its recommendation on "Language Resources for the Future – The Future of Language Resources, The Strategic Language Resource Agenda" is an essential roadmap for us.

One of the conclusions that has been thoroughly debated within the board of ELRA is the set-up of a new permanent forum, gathering all LREC attendees and all interested individuals to constitute the Language Resources and Evaluation Forum (LRE-F). We feel that it is important to identify and gather the members of this very broad community and ensure that interactive exchanges/services can be set up to help them work together. The forum is established at this LREC 2012 where the largest group of individuals that have to do with Language Resources and Evaluation are present; it is open (and not limited) to: scientists, students or professors, involved in research activities in universities, small and medium companies or international groups; decision-makers or project managers in large public institutions, etc. You have been invited to join when registering for LREC and we hope you expressed your wish to join. Those who missed that opportunity still can do so at any time through the ELRA portal. Among the services, members of the LRE-F will be offered free downloading of many resources from the ELRA Catalogue and the META-SHARE repository, access to the legal helpdesk, access to the LRE Map, the LR Library, access to LRE Wiki, etc. Members of the community will be also encouraged to join so to upload resources on the ELRA and/or ELRA-META-SHARE repository to share with other colleagues.

An additional service offered by ELRA to all its partners, is the production, customization, repurposing of Language Resources, on demand. ELRA, through the ELDA staff, is involved in LR production. Such productions comprised speech corpora, lexica, textual corpora, both monolingual and aligned / comparable multilingual ones, video and audio data, documents and many other modalities. Such activities included production from scratch as well as, repurposing of existing ones, merging of various sets, annotations, transcriptions, META-DATA labeling of existing databases, etc. ELRA carried out such production for more than 30 languages, working proudly with hundreds of local partners all over the world.

In order to turn this into efficient and cost-effective services, ELDA is part of the EC project PANACEA (Platform for Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources). The project aimed at building a factory of Language Resources that "*progressively automates the stages involved in the acquisition, production, updating and maintenance of language resources*", in particular those required by MT systems. The platform will be available both as a Framework (software package) for partners to deploy and as a service offered by ELRA for specific production of resources.

ELRA is ready to assist in LR productions, at any of the needed stages.

ELRA continues to produce resources for technology evaluation and the related campaigns. We would like to stress the importance of packaging the LR and methodologies used for such purposes, to help other interested colleagues in carrying similar assessments. It is also crucial to review such resources for possible repurposing for other needs. ELRA is prepared to assist all evaluators in these tasks. More than 50 packages are already available through ELRA catalogue, most of them for free. In order to keep an efficient stream of information on this, ELRA continues to support the HLT evaluation portal (www.hlt-evaluation.org).

While preparing this message, I went back to messages of our first gathering in Granada (LREC'1998), ages ago one would think!

> *"The presently embryonic infrastructure should be reinforced, so that the same infrastructure is able to coordinate and perform, avoiding duplications, different complementary tasks: to provide and update the general repertories of linguistic data and knowledge which should be available for as many languages as possible, to produce at reasonable costs and in due time customized LR to answer specific requests of developers, to offer services the community urgently needs, information, consultation, validation, etc. "* (Antonio Zampolli, Introductory message to LREC 1998, Granada)

After the set-up and consolidation of ELRA, and now with our strong commitment to boost and sustain META-SHARE, we feel these new approaches to efficient and cost effective sharing of LRs are essential milestones for our community and ELRA is very proud to play a role in this effort.

Last but not least, let me tell you a few words about our week here in Istanbul. In addition to the technical and scientific program (see more details in our LREC Chair message, herein), we have designed, with our local colleagues, a social program to make our stay enjoyable but also fruitful for establishing new relationships and networks, setting up new projects and collaborations, and above all making new friends.

We did our best to make your stay in Istanbul a very pleasant experience, we hope that both our welcome reception (Wednesday, May 23) and Gala Dinner (Friday, 25 may) will give you memories to treasure. We hope that during these events and throughout the week, we will show you some of the best Turkey as to offer.

As always, we tried to introduce novelties and new features to improve the organization of LREC.

In addition to the EU Village, a dissemination / exhibition opportunity for EU projects, we have extended this with an EU "track" of oral presentations, to offer you a full afternoon of information on the major activities supported by the EC (Thursday, May 24).

LREC 2012 will definitely close the chapter of proceedings supplied as hardcopies, CDs or USBs. We will keep the tradition to provide the participants with hardcopies of the program booklet, and the abstracts (of papers of the main conference and the workshops, material of tutorials). BUT the proceedings will only be made available and in advance, on the LREC web site, and in various format, so that you can download them on your favorite media and bring them with you. Please do that in advance, local Internet connection may not be efficient enough for all of us to do that locally.

A new experiment will be conducted this time, a tool, called MyLREC-program, will allow participants to choose their sessions (even the papers they would like to hear within a given session), design their own program and plan their days. One can print it as a PDF file or import it in one's favorite calendar. Please visit the LREC2012 pages for this. We hope this will help you navigate efficiently and friendly through all the sessions LREC is offering.

Finally, I wish to express my deep thanks to our partners and supporters, who throughout the years make LREC so successful.

I would like first to thank our Silver Sponsors: CELI, NUANCE; our bronze sponsors: EML (The European Media Laboratory GmbH), IMMI, K-dictionaries, META-NET, and Quaero.

I would like also to thank the EC Village participants; we hope that such gathering will offer them an opportunity to foster their dissemination and hopefully discuss exploitation plans with the attendees.

I would like to thank the LREC Local Committee, chaired by Mehmed Özkan, who helped us with all logistic issues.

I would like finally to warmly thank the joint team of the two institutions that devote 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 Sara Goggi and Hélène Mazo and the team: Victoria Arranz, Cécile Barbier, Paola Baroni, Roberto Bartolini, Riccardo Del Gratta, Francesca Frontini, Olivier Hamon, Valerie Mapelli, Vincenzo Parrinelli, Valeria Quochi, Caroline Rannaud, Irene Russo, Priscille Schneller.

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

Once again, welcome to Istanbul, welcome to LREC' 2012

Khalid Choukri
ELRA Secretary General and ELDA Managing Director

Message of the Chair of the Local Organizing Committee
Mehmed Özkan

Language; other than being a means of communicating human thoughts and feelings in an everyday life, is also a barrier separating civilizations apart and defining borders among the nations and peoples. Through out the history, clusters formed around languages have long provided fertile yet different habitats of cultural and social islands, enabling the cultivation of the human experience and knowledge in a relatively exclusive manner. In a way this natural isolation of human clusters made the concurrent exploration of the knowledge possible, leading to a synergetic conquest of understanding the individual's own existence. For the expected synergy to surface however, the proper means of interaction must also be present. In the old times and still for many cases the best channels of cross-cultural communication have been the multilingual humans and their products. If the language barrier is a one dominant reason for the existence of fertile cultural islands then we can argue these barriers must continue their presence, despite the transformation towards "globalization". On the other hand, for the humankind to benefit from own achievements and carry on to the next generations with advancements, the islands of knowledge are better served when shared.

With the electronic age and its sub-stages the storage and flow of knowledge reached to a point, possibly beyond the imagination of the inventors of the information technologies, and it is growing faster than ever. In the beginning the default language to represent this massive knowledge was English. Gradually came the others. Already, it is estimated less than 20% of the information in the cyberspace is in English and decreasing. Once again the languages are claiming the borders and once again we need translators, interpreters to benefit from this vast human knowledge. This time however, the tools we need are not only for understanding the knowledge in other languages, but also to coop with the massive size and enormous speed of the information we are faced with even in our own mother tongue. LREC is playing a noble role for achieving this goal by bringing the most important resources together to tackle the problem, that is the scientist and engineers committed to develop the much needed language tools and resources.

Istanbul, located in the crossroads of civilizations, continents and important seas have lived the benefits of exchanging knowledge and prospered by attracting the scientists, artists, architects, poets, philosophers and writers from all around the world throughout the history. Named as capitals of several empires and a sultanate, was latest the European Capital of Culture in 2010. She is indeed and has been the Capital of hearts for many. For some it was the crossing of the Silk Road, for some other the ultimate spice outpost, or the final destination of the Orient Express. However we perceive it, Istanbul has been a merger point of architectural details of three continents; a meeting point of religions and of courses the languages. She lived it all… Once again it is a pleasure to see Istanbul bringing the respectable scientific community who are committed to bring people and knowledge together with innovative language and speech technologies. I am honored for being your host during this prestigious event and welcome you.

Hoping your LREC 2012 experience in Istanbul will be memorable one that you will always remember with a smile …

Mehmed Özkan
Chair of the Local Committee

Table of Contents

O1 - Corpora for Machine Translation

Wednesday, May 23, 11:35

Chairperson: **Josef van Genabith** Oral Session

PaCo2: A Fully Automated tool for gathering Parallel Corpora from the Web

Iñaki San Vicente and Iker Manterola

The importance of parallel corpora in the NLP field is fully acknowledged. This paper presents a tool that can build parallel corpora given just a seed word list and a pair of languages. Our approach is similar to others proposed in the literature, but introduces a new phase to the process. While most of the systems leave the task of finding websites containing parallel content up to the user, PaCo2 (Parallel Corpora Collector) takes care of that as well. The tool is language independent as far as possible, and adapting the system to work with new languages is fairly straightforward. Evaluation of the different modules has been carried out for Basque-Spanish, Spanish-English and Portuguese-English language pairs. Even though there is still room for improvement, results are positive. Results show that the corpora created have very good quality translations units, and the quality is maintained for the various language pairs. Details of the corpora created up until now are also provided.

Terra: a Collection of Translation Error-Annotated Corpora

Mark Fishel, Ondřej Bojar and Maja Popović

Recently the first methods of automatic diagnostics of machine translation have emerged; since this area of research is relatively young, the efforts are not coordinated. We present a collection of translation error-annotated corpora, consisting of automatically produced translations and their detailed manual translation error analysis. Using the collected corpora we evaluate the available state-of-the-art methods of MT diagnostics and assess, how well the methods perform, how they compare to each other and whether they can be useful in practice.

A light way to collect comparable corpora from the Web

Ahmet Aker, Evangelos Kanoulas and Robert Gaizauskas

Statistical Machine Translation (SMT) relies on the availability of rich parallel corpora. However, in the case of under-resourced languages, parallel corpora are not readily available. To overcome this problem previous work has recognized the potential of using comparable corpora as training data. The process of obtaining such data usually involves (1) downloading a separate list of documents for each language, (2) matching the documents between two languages usually by comparing the document contents, and finally (3) extracting useful data for SMT from the matched document pairs. This process requires a large amount of time and resources since a huge volume of documents needs to be downloaded to increase the chances of finding good document pairs. In this work we aim to reduce the amount of time and resources spent for tasks 1 and 2. Instead of obtaining full documents we first obtain just titles along with some meta-data such as time and date of publication. Titles can be obtained through Web Search and RSS News feed collections so that download of the full documents is not needed. We show experimentally that titles can be used to approximate the comparison between documents using full document contents.

SUMAT: Data Collection and Parallel Corpus Compilation for Machine Translation of Subtitles

Volha Petukhova, Rodrigo Agerri, Mark Fishel, Sergio Penkale, Arantza del Pozo, Mirjam Sepesy Maucec, Andy Way, Panayota Georgakopoulou and Martin Volk

Subtitling and audiovisual translation have been recognized as areas that could greatly benefit from the introduction of Statistical Machine Translation (SMT) followed by post-editing, in order to increase efficiency of subtitle production process. The FP7 European project SUMAT (An Online Service for SUbtitling by MAchine Translation: http://www.sumat-project.eu) aims to develop an online subtitle translation service for nine European languages, combined into 14 different language pairs, in order to semi-automate the subtitle translation processes of both freelance translators and subtitling companies on a large scale. In this paper we discuss the data collection and parallel corpus compilation for training SMT systems, which includes several procedures such as data partition, conversion, formatting, normalization and alignment. We discuss in detail each data pre-processing step using various approaches. Apart from the quantity (around 1 million subtitles per language pair), the SUMAT corpus has a number of very important characteristics. First of all, high quality both in terms of translation and in terms of high-precision alignment of parallel documents and their contents has been achieved. Secondly, the contents are provided in one consistent format and encoding. Finally, additional information such as type of content in terms of genres and domain is available.

The FAUST Corpus of Adequacy Assessments for Real-World Machine Translation Output

Daniele Pighin, Lluís Màrquez and Lluís Formiga

We present a corpus consisting of 11,292 real-world English to Spanish automatic translations annotated with relative (ranking) and absolute (adequate/non-adequate) quality assessments. The

translation requests, collected through the popular translation portal http://reverso.net, provide a most variated sample of real-world machine translation (MT) usage, from complete sentences to units of one or two words, from well-formed to hardly intelligible texts, from technical documents to colloquial and slang snippets. In this paper, we present 1) a preliminary annotation experiment that we carried out to select the most appropriate quality criterion to be used for these data, 2) a graph-based methodology inspired by Interactive Genetic Algorithms to reduce the annotation effort, and 3) the outcomes of the full-scale annotation experiment, which result in a valuable and original resource for the analysis and characterization of MT-output quality.

O2 - Infrastructures and Strategies for LRs (1)

Wednesday, May 23, 11:35

Chairperson: **Hans Uszkoreit** Oral Session

The META-SHARE Language Resources Sharing Infrastructure: Principles, Challenges, Solutions

Stelios Piperidis

Language resources have become a key factor in the development cycle of language technology. The current prevailing methodologies, the sheer number of languages and the vast volumes of digital content together with the wide palette of useful content processing applications, render new models for managing the underlying language resources indispensable. This paper presents META-SHARE, an open resource exchange infrastructure, which aims to boost visibility, documentation, identification, openness and sharing, collaboration, preservation and interoperability of language data and basic language processing tools. META-SHARE is implemented as a network of distributed repositories of language resources. It offers providers and consumers of resources the necessary functionalities for describing, storing, searching, licensing and downloading language resources in a single integrated technical platform. META-SHARE favours and aligns itself with the growing open data and open source tools movement. To this end, it has prepared the necessary underlying legal framework consisting of a Charter for language resource sharing, as well as a set of licensing templates aiming to act as recommended licence models in an attempt to facilitate the legal interoperability of language resources. In its current version, META-SHARE features 13 resource repositories, with over 1200 resource packages.

The Language Library: supporting community effort for collective resource production

Riccardo Del Gratta, Francesca Frontini, Francesco Rubino, Irene Russo and Nicoletta Calzolari

Relations among phenomena at different linguistic levels are at the essence of language properties but today we focus mostly on one specific linguistic layer at a time, without (having the possibility of) paying attention to the relations among the different layers. At the same time our efforts are too much scattered without much possibility of exploiting other people's achievements. To address the complexities hidden in multilayer interrelations even small amounts of processed data can be useful, improving the performance of complex systems. Exploiting the current trend towards sharing we want to initiate a collective movement that works towards creating synergies and harmonisation among different annotation efforts that are now dispersed. In this paper we present the general architecture of the Language Library, an initiative which is conceived as a facility for gathering and making available through simple functionalities the linguistic knowledge the field is able to produce, putting in place new ways of collaboration within the LRT community. In order to reach this goal, a first population round of the Language Library has started around a core of parallel/comparable texts that have been annotated by several contributors submitting a paper for LREC2012. The Language Library has also an ancillary aim related to language documentation and archiving and it is conceived as a theory-neutral space which allows for several language processing philosophies to coexist.

Using the International Standard Language Resource Number: Practical and Technical Aspects

Jungyeul Park, Victoria Arranz, Olivier Hamon and Khalid Choukri

This paper describes the International Standard Language Resource Number (ISLRN), a new identification schema for Language Resources where a Language Resource is provided with a unique and universal name using a standardized nomenclature. This will ensure that Language Resources be identified, accessed and disseminated in a unique manner, thus allowing them to be recognized with proper references in all activities concerning Human Language Technologies as well as in all documents and scientific papers. This would allow, for instance, the formal identification of potentially repeated resources across different repositories, the formal referencing of language resources and their correct use when different versions are processed by tools.

ELRA in the heart of a cooperative HLT world

Valérie Mapelli, Victoria Arranz, Matthieu Carré, Hélène Mazo, Djamel Mostefa and Khalid Choukri

This paper aims at giving an overview of ELRA's recent activities. The first part elaborates on ELRA's means of boosting the sharing Language Resources (LRs) within the HLT community through its catalogues, LRE-Map initiative, as well as its work towards the integration of its LRs within the META-SHARE open infrastructure. The second part shows how ELRA helps in the development and evaluation of HLT, in particular through its numerous participations to collaborative projects for the production of resources and platforms to facilitate their production and exploitation. A third part focuses on ELRA's work for clearing IPR issues in a HLT-oriented context, one of its latest initiative being its involvement in a Fair Research Act proposal to promote the easy access to LRs to the widest community. Finally, the last part elaborates on recent actions for disseminating information and promoting cooperation in the field, e.g. an the Language Library being launched at LREC2012 and the creation of an International Standard LR Number, a LR unique identifier to enable the accurate identification of LRs. Among the other messages ELRA will be conveying the attendees are the announcement of a set of freely available resources, the establishment of a LR and Evaluation forum, etc.

Twenty Years of Language Resource Development and Distribution: A Progress Report on LDC Activities

Christopher Cieri, Marian Reed, Denise DiPersio and Mark Liberman

On the Linguistic Data Consortium's (LDC) 20th anniversary, this paper describes the changes to the language resource landscape over the past two decades, how LDC has adjusted its practice to adapt to them and how the business model continues to grow. Specifically, we will discuss LDC's evolving roles and changes in the sizes and types of LDC language resources (LR) as well as the data they include and the annotations of that data. We will also discuss adaptations of the LDC business model and the sponsored projects it supports.

O3 - Semantics

Wednesday, May 23, 11:35

Chairperson: **Bolette Pedersen** Oral Session

Polaris: Lymba's Semantic Parser

Dan Moldovan and Eduardo Blanco

Semantic representation of text is key to text understanding and reasoning. In this paper, we present Polaris, Lymba's semantic parser. Polaris is a supervised semantic parser that given text extracts semantic relations. It extracts relations from a wide variety of lexico-syntactic patterns, including verb-argument structures, noun compounds and others. The output can be provided in several formats: XML, RDF triples, logic forms or plain text, facilitating interoperability with other tools. Polaris is implemented using eight separate modules. Each module is explained and a detailed example of processing using a sample sentence is provided. Overall results using a benchmark are discussed. Per module performance, including errors made and pruned by each module are also analyzed.

Automatic classification of German"an"particle verbs

Sylvia Springorum, Sabine Schulte im Walde and Antje Roßdeutscher

The current study works at the interface of theoretical and computational linguistics to explore the semantic properties of an particle verbs, i.e., German particle verbs with the particle an. Based on a thorough analysis of the particle verbs from a theoretical point of view, we identified empirical features and performed an automatic semantic classification. A focus of the study was on the mutual profit of theoretical and empirical perspectives with respect to salient semantic properties of the an particle verbs: (a) how can we transform the theoretical insights into empirical, corpus-based features, (b) to what extent can we replicate the theoretical classification by a machine learning approach, and (c) can the computational analysis in turn deepen our insights to the semantic properties of the particle verbs? The best classification result of 70% correct class assignments was reached through a GermaNet-based generalization of direct object nouns plus a prepositional phrase feature. These particle verb features in combination with a detailed analysis of the results at the same time confirmed and enlarged our knowledge about salient properties.

Pragmatic identification of the witness sets

Livio Robaldo and Jakub Szymanik

Among the readings available for NL sentences, those where two or more sets of entities are independent of one another are particularly challenging from both a theoretical and an empirical point of view. Those readings are termed here as 'Independent Set (IS) readings'. Standard examples of such readings are the well-known Collective and Cumulative Readings. (Robaldo, 2011) proposes a logical framework that can properly represent the meaning of IS readings in terms of a set-Skolemization of the witness sets. One of the main assumptions of Robaldo's logical framework, drawn from (Schwarzschild, 1996), is that pragmatics

plays a crucial role in the identification of such witness sets. Those are firstly identified on pragmatic grounds, then logical clauses are asserted on them in order to trigger the appropriate inferences. In this paper, we present the results of an experimental analysis that appears to confirm Robaldo's hypotheses concerning the pragmatic identification of the witness sets.

Evaluating automatic cross-domain Dutch semantic role annotation

Orphée De Clercq, Veronique Hoste and Paola Monachesi

In this paper we present the first corpus where one million Dutch words from a variety of text genres have been annotated with semantic roles. 500K have been completely manually verified and used as training material to automatically label another 500K. All data has been annotated following an adapted version of the PropBank guidelines. The corpus's rich text type diversity and the availability of manually verified syntactic dependency structures allowed us to experiment with an existing semantic role labeler for Dutch. In order to test the system's portability across various domains, we experimented with training on individual domains and compared this with training on multiple domains by adding more data. Our results show that training on large data sets is necessary but that including genre-specific training material is also crucial to optimize classification. We observed that a small amount of in-domain training data is already sufficient to improve our semantic role labeler.

Logic Based Methods for Terminological Assessment

Benoît Robichaud

We present a new version of a Graphical User Interface (GUI) called DiCoInfo Visuel, mainly based on a graph visualization device and used for exploring and assessing lexical data found in DiCoInfo, a specialized e-dictionary of computing and the Internet. This new GUI version takes advantage of the fundamental nature of the lexical network encoded in the dictionary: it uses logic based methods from logic programming to explore relations between entries and find pieces of relevant information that may be not accessible by direct searches. The result is a more realistic and useful data coverage shown to end users.

O4 - Speech corpora

Wednesday, May 23, 11:35

Chairperson: **Sophie Rosset** Oral Session

KALAKA-2: a TV Broadcast Speech Database for the Recognition of Iberian Languages in Clean and Noisy Environments

Luis Javier Rodriguez-Fuentes, Mikel Penagarikano, Amparo Varona, Mireia Diez and German Bordel

This paper presents the main features (design issues, recording setup, etc.) of KALAKA-2, a TV broadcast speech database specifically designed for the development and evaluation of language recognition systems in clean and noisy environments. KALAKA-2 was created to support the Albayzin 2010 Language Recognition Evaluation (LRE), organized by the Spanish Network on Speech Technologies from June to November 2010. The database features 6 target languages: Basque, Catalan, English, Galician, Portuguese and Spanish, and includes segments in other (Out-Of-Set) languages, which allow to perform open-set verification tests. The best performance attained in the Albayzin 2010 LRE is presented and briefly discussed. The performance of a state-of-the-art system in various tasks defined on the database is also presented. In both cases, results highlight the suitability of KALAKA-2 as a benchmark for the development and evaluation of language recognition technology.

The C-ORAL-BRASIL I: Reference Corpus for Spoken Brazilian Portuguese

Tommaso Raso, Heliana Mello and Maryualê Malvessi Mittmann

C-ORAL-BRASIL I is a Brazilian Portuguese spontaneous speech corpus compiled following the same architecture adopted by the C-ORAL-ROM resource. The main goal is the documentation of the diaphasic and diastratic variations in Brazilian Portuguese. The diatopic variety represented is that of the metropolitan area of Belo Horizonte, capital city of Minas Gerais. Even though it was not a primary goal, a nice balance was achieved in terms of speakers' diastratic features (sex, age and school level). The corpus is entirely dedicated to informal spontaneous speech and comprises 139 informal speech texts, 208,130 words and 21:08:52 hours of recording, distributed into family/private (80%) and public (20%) contexts. The LR includes audio files, transcripts in text format and text-to-speech alignment (accessible with WinPitch Pro software). C-ORAL-BRASIL I also provides transcripts with Part-of-Speech annotation implemented through the parser system Palavras. Transcripts were validated regarding the proper application of transcription criteria and also for the

annotation of prosodic boundaries. Some quantitative features of C-ORAL-BRASIL I in comparison with the informal C-ORAL-ROM are reported.

The ETAPE corpus for the evaluation of speech-based TV content processing in the French language

Guillaume Gravier, Gilles Adda, Niklas Paulsson, Matthieu Carré, Aude Giraudel and Olivier Galibert

The paper presents a comprehensive overview of existing data for the evaluation of spoken content processing in a multimedia framework for the French language. We focus on the ETAPE corpus which will be made publicly available by ELDA mid 2012, after completion of the evaluation campaign, and recall existing resources resulting from previous evaluation campaigns. The ETAPE corpus consists of 30 hours of TV and radio broadcasts, selected to cover a wide variety of topics and speaking styles, emphasizing spontaneous speech and multiple speaker areas.

Automatic Speech Recognition on a Firefighter TETRA Broadcast Channel

Daniel Stein and Bela Usabaev

For a reliable keyword extraction on firefighter radio communication, a strong automatic speech recognition system is needed. However, real-life data poses several challenges like a distorted voice signal, background noise and several different speakers. Moreover, the domain is out-of-scope for common language models, and the available data is scarce. In this paper, we introduce the PRONTO corpus, which consists of German firefighter exercise transcriptions. We show that by standard adaption techniques the recognition rate already rises from virtually zero to up to 51.7% and can be further improved by domain-specific rules to 47.9%. Extending the acoustic material by semi-automatic transcription and crawled in-domain written material, we arrive at a WER of 45.2%.

TED-LIUM: an Automatic Speech Recognition dedicated corpus

Anthony Rousseau, Paul Deléglise and Yannick Estève

This paper presents the corpus developed by the LIUM for Automatic Speech Recognition (ASR), based on the TED Talks. This corpus was built during the IWSLT 2011 Evaluation Campaign, and is composed of 118 hours of speech with its accompanying automatically aligned transcripts. We describe the content of the corpus, how the data was collected and processed, how it will be publicly available and how we built an ASR system using this data leading to a WER score of 17.4 %. The official results we obtained at the IWSLT 2011 evaluation campaign are also discussed.

P1 - Anaphora and Coreference

Wednesday, May 23, 11:35

Chairperson: **Ineke Schuurman** Poster Session

QurAna: Corpus of the Quran annotated with Pronominal Anaphora

Abdul-Baquee Sharaf and Eric Atwell

This paper presents QurAna: a large corpus created from the original Quranic text, where personal pronouns are tagged with their antecedence. These antecedents are maintained as an ontological list of concepts, which have proved helpful for information retrieval tasks. QurAna is characterized by: (a) comparatively large number of pronouns tagged with antecedent information (over 24,500 pronouns), and (b) maintenance of an ontological concept list out of these antecedents. We have shown useful applications of this corpus. This corpus is first of its kind considering classical Arabic text, which could be used for interesting applications for Modern Standard Arabic as well. This corpus would benefit researchers in obtaining empirical and rules in building new anaphora resolution approaches. Also, such corpus would be used to train, optimize and evaluate existing approaches.

The Use of Parallel and Comparable Data for Analysis of Abstract Anaphora in German and English

Stefanie Dipper, Melanie Seiss and Heike Zinsmeister

Parallel corpora — original texts aligned with their translations — are a widely used resource in computational linguistics. Translation studies have shown that translated texts often differ systematically from comparable original texts. Translators tend to be faithful to structures of the original texts, resulting in a“shining through”of the original language preferences in the translated text. Translators also tend to make their translations most comprehensible with the effect that translated texts can be more explicit than their source texts. Motivated by the need to use a parallel resource for cross-linguistic feature induction in abstract anaphora resolution, this paper investigates properties of English and German texts in the Europarl corpus, taking into account both general features such as sentence length as well as task-dependent features such as the distribution of demonstrative noun phrases. The investigation is based on the entire Europarl corpus as well as on a small subset thereof, which has been manually annotated. The results indicate English translated texts

are sufficiently "authentic" to be used as training data for anaphora resolution; results for German texts are less conclusive, though.

Interplay of Coreference and Discourse Relations: Discourse Connectives with a Referential Component

Lucie Poláková, Pavlína Jínová and Jiří Mírovský

This contribution explores the subgroup of text structuring expressions with the form preposition + demonstrative pronoun, thus it is devoted to an aspect of the interaction of coreference relations and relations signaled by discourse connectives (DCs) in a text. The demonstrative pronoun typically signals a referential link to an antecedent, whereas the whole expression can, but does not have to, carry a discourse meaning in sense of discourse connectives. We describe the properties of these phrases/expressions with regard to their antecedents, their position among the text-structuring language means and their features typical for the "connective function" of them compared to their "non-connective function". The analysis is carried out on Czech data from the approx. 50,000 sentences of the Prague Dependency Treebank 2.0, directly on the syntactic trees. We explore the characteristics of these phrases/expressions discovered during two projects: the manual annotation of 1, coreference relations (Nedoluzhko et al. 2011) and 2, discourse connectives, their scopes and meanings (Mladová et al. 2008).

A Portuguese-Spanish Corpus Annotated for Subject Realization and Referentiality

Luz Rello and Iria Gayo

This paper presents a comparable corpus of Portuguese and Spanish consisting of legal and health texts. We describe the annotation of zero subject, impersonal constructions and explicit subjects in the corpus. We annotated 12,492 examples using a scheme that distinguishes between different linguistic levels (phonology, syntax, semantics, etc.) and present a taxonomy of instances on which annotators disagree. The high level of inter-annotator agreement (83%-95%) and the performance of learning algorithms trained on the corpus show that our corpus is a reliable and useful resource.

Coreference in Spoken vs. Written Texts: a Corpus-based Analysis

Marilisa Amoia, Kerstin Kunz and Ekaterina Lapshinova-Koltunski

This paper describes an empirical study of coreference in spoken vs. written text. We focus on the comparison of two particular text types, interviews and popular science texts, as instances of spoken and written texts since they display quite different discourse structures. We believe in fact, that the correlation of difficulties in coreference resolution and varying discourse structures requires a deeper analysis that accounts for the diversity of coreference strategies or their sub-phenomena as indicators of text type or genre. In this work, we therefore aim at defining specific parameters that classify differences in genres of spoken and written texts such as the preferred segmentation strategy, the maximal allowed distance in or the length and size of coreference chains as well as the correlation of structural and syntactic features of coreferring expressions. We argue that a characterization of such genre dependent parameters might improve the performance of current state-of-art coreference resolution technology.

Annotating Near-Identity from Coreference Disagreements

Marta Recasens, M. Antònia Martí and Constantin Orasan

We present an extension of the coreference annotation in the English NP4E and the Catalan AnCora-CA corpora with near-identity relations, which are borderline cases of coreference. The annotated subcorpora have 50K tokens each. Near-identity relations, as presented by Recasens et al. (2010; 2011), build upon the idea that identity is a continuum rather than an either/or relation, thus introducing a middle ground category to explain currently problematic cases. The first annotation effort that we describe shows that it is not possible to annotate near-identity explicitly because subjects are not fully aware of it. Therefore, our second annotation effort used an indirect method, and arrived at near-identity annotations by inference from the disagreements between five annotators who had only a two-alternative choice between coreference and non-coreference. The results show that whereas as little as 2-6% of the relations were explicitly annotated as near-identity in the former effort, up to 12-16% of the relations turned out to be near-identical following the indirect method of the latter effort.

This also affects the context - Errors in extraction based summaries

Thomas Kaspersson, Christian Smith, Henrik Danielsson and Arne Jönsson

Although previous studies have shown that errors occur in texts summarized by extraction based summarizers, no study has investigated how common different types of errors are and how that changes with degree of summarization. We have conducted studies of errors in extraction based single document summaries using 30 texts, summarized to 5 different degrees and tagged for errors by human judges. The results show that the most common errors are absent cohesion or context and various types

of broken or missing anaphoric references. The amount of errors is dependent on the degree of summarization where some error types have a linear relation to the degree of summarization and others have U-shaped or cut-off linear relations. These results show that the degree of summarization has to be taken into account to minimize the amount of errors by extraction based summarizers.

Annotation of anaphoric relations and topic continuity in Japanese conversation

Natsuko Nakagawa and Yasuharu Den

This paper proposes a basic scheme for annotating anaphoric relations in Japanese conversations. More specifically, we propose methods of (i) dividing discourse segments into meaningful units, (ii) identifying zero pronouns and other overt anaphors, (iii) classifying zero pronouns, and (iv) identifying anaphoric relations. We discuss various kinds of problems involved in the annotation mainly caused by on-line processing of discourse and/or interactions between the participants. These problems do not arise in annotating written languages. This paper also proposes a method to compute topic continuity based on anaphoric relations. The topic continuity involves the information status of the noun in question (given, accessible, and new) and persistence (whether the noun is mentioned multiple times or not). We show that the topic continuity correlates with short-utterance units, which are determined prosodically through the previous annotations; nouns of high topic continuity tend to be prosodically separated from the predicates. This result indicates the validity of our annotations of anaphoric relations and topic continuity and the usefulness for further studies on discourse and interaction.

Domain-specific vs. Uniform Modeling for Coreference Resolution

Olga Uryupina and Massimo Poesio

Several corpora annotated for coreference have been made available in the past decade. These resources differ with respect to their size and the underlying structure: the number of domains and their similarity. Our study compares domain-specific models, learned from small heterogeneous subsets of the investigated corpora, against uniform models, that utilize all the available data. We show that for knowledge-poor baseline systems, domain-specific and uniform modeling yield same results. Systems, relying on large amounts of linguistic knowledge, however, exhibit differences in their performance: with all the designed features in use, domain-specific models suffer from over-fitting, whereas with pre-selected feature sets they tend to outperform union models.

Creating a Coreference Resolution System for Polish

Mateusz Kopeć and Maciej Ogrodniczuk

Although the availability of the natural language processing tools and the development of metrics to evaluate them increases, there is a certain gap to fill in that field for the less-resourced languages, such as Polish. Therefore the projects which are designed to extend the existing tools for diverse languages are the best starting point for making these languages more and more covered. This paper presents the results of the first attempt of the coreference resolution for Polish using statistical methods. It presents the conclusions from the process of adapting the Beautiful Anaphora Resolution Toolkit (BART; a system primarily designed for the English language) for Polish and collates its evaluation results with those of the previously implemented rule-based system. Finally, we describe our plans for the future usage of the tool and highlight the upcoming research to be conducted, such as the experiments of a larger scale and the comparison with other machine learning tools.

P2 - Tools, Systems and Evaluation

Wednesday, May 23, 11:35

Chairperson: **Michael Kipp** Poster Session

Fast Labeling and Transcription with the Speechalyzer Toolkit

Felix Burkhardt

We describe a software tool named “Speechalyzer” which is optimized to process large speech data sets with respect to transcription, labeling and annotation. It is implemented as a client server based framework in Java and interfaces software for speech recognition, synthesis, speech classification and quality evaluation. The application is mainly the processing of training data for speech recognition and classification models and performing benchmarking tests on speech to text, text to speech and speech categorization software systems.

Automatic annotation of head velocity and acceleration in Anvil

Bart Jongejan

We describe an automatic face tracker plugin for the ANVIL annotation tool. The face tracker produces data for velocity and for acceleration in two dimensions. We compare annotations generated by the face tracking algorithm with independently made manual annotations for head movements. The annotations are a useful supplement to manual annotations and may help human annotators to quickly and reliably determine onset of head

movements and to suggest which kind of head movement is taking place.

AVATecH – automated annotation through audio and video analysis

Przemyslaw Lenkiewicz, Binyam Gebrekidan Gebre, Oliver Schreer, Stefano Masneri, Daniel Schneider and Sebastian Tschöpel

In different fields of the humanities annotations of multimodal resources are a necessary component of the research workflow. Examples include linguistics, psychology, anthropology, etc. However, creation of those annotations is a very laborious task, which can take 50 to 100 times the length of the annotated media, or more. This can be significantly improved by applying innovative audio and video processing algorithms, which analyze the recordings and provide automated annotations. This is the aim of the AVATecH project, which is a collaboration of the Max Planck Institute for Psycholinguistics (MPI) and the Fraunhofer institutes HHI and IAIS. In this paper we present a set of results of automated annotation together with an evaluation of their quality.

An Oral History Annotation Tool for INTER-VIEWs

Henk van den Heuvel, Eric Sanders, Robin Rutten, Stef Scagliola and Paula Witkamp

We present a web-based tool for retrieving and annotating audio fragments of e.g. interviews. Our collection contains 250 interviews with veterans of Dutch conflicts and military missions. The audio files of the interviews were disclosed using ASR technology focussed at keyword retrieval. Resulting transcripts were stored in a MySQL database together with metadata, summary texts, and keywords, and carefully indexed. Retrieved fragments can be made audible and annotated. Annotations can be kept personal or be shared with other users. The tool and formats comply with CLARIN standards. A demo version of the tool is available at http://wwwlands2.let.kun.nl/spex/annotationtooldemo.

ELAN development, keeping pace with communities' needs

Han Sloetjes and Aarthy Somasundaram

ELAN is a versatile multimedia annotation tool that is being developed at the Max Planck Institute for Psycholinguistics. About a decade ago it emerged out of a number of corpus tools and utilities and it has been extended ever since. This paper focuses on the efforts made to ensure that the application keeps up with the growing needs of that era in linguistics and multimodality research; growing needs in terms of length and resolution of recordings, the number of recordings made and transcribed and the number of levels of annotation per transcription.

Inforex – a web-based tool for text corpus management and semantic annotation

Michał Marcińczuk, Jan Kocoń and Bartosz Broda

The aim of this paper is to present a system for semantic text annotation called Inforex. Inforex is a web-based system designed for managing and annotating text corpora on the semantic level including annotation of Named Entities (NE), anaphora, Word Sense Disambiguation (WSD) and relations between named entities. The system also supports manual text clean-up and automatic text pre-processing including text segmentation, morphosyntactic analysis and word selection for word sense annotation. Inforex can be accessed from any standard-compliant web browser supporting JavaScript. The user interface has a form of dynamic HTML pages using the AJAX technology. The server part of the system is written in PHP and the data is stored in MySQL database. The system make use of some external tools that are installed on the server or can be accessed via web services. The documents are stored in the database in the original format — either plain text, XML or HTML. Tokenization and sentence segmentation is optional and is stored in a separate table. Tokens are stored as pairs of values representing indexes of first and last character of the tokens and sets of features representing the morpho-syntactic information.

Towards Automatic Gesture Stroke Detection

Binyam Gebrekidan Gebre, Peter Wittenburg and Przemyslaw Lenkiewicz

Automatic annotation of gesture strokes is important for many gesture and sign language researchers. The unpredictable diversity of human gestures and video recording conditions require that we adopt a more adaptive case-by-case annotation model. In this paper, we present a work-in progress annotation model that allows a user to a) track hands/face b) extract features c) distinguish strokes from non-strokes. The hands/face tracking is done with color matching algorithms and is initialized by the user. The initialization process is supported with immediate visual feedback. Sliders are also provided to support a user-friendly adjustment of skin color ranges. After successful initialization, features related to positions, orientations and speeds of tracked hands/face are extracted using unique identifiable features (corners) from a window of frames and are used for training a learning algorithm. Our preliminary results for stroke detection under non-ideal video conditions are promising and show the potential applicability of our methodology.

EXMARaLDA and the FOLK tools – two toolsets for transcribing and annotating spoken language

Thomas Schmidt

This paper presents two toolsets for transcribing and annotating spoken language: the EXMARaLDA system, developed at the University of Hamburg, and the FOLK tools, developed at the Institute for the German Language in Mannheim. Both systems are targeted at users interested in the analysis of spontaneous, multi-party discourse. Their main user community is situated in conversation analysis, pragmatics, sociolinguistics and related fields. The paper gives an overview of the individual tools of the two systems – the Partitur-Editor, a tool for multi-level annotation of audio or video recordings, the Corpus Manager, a tool for creating and administering corpus metadata, EXAKT, a query and analysis tool for spoken language corpora, FOLKER, a transcription editor optimized for speed and efficiency of transcription, and OrthoNormal, a tool for orthographical normalization of transcription data. It concludes with some thoughts about the integration of these tools into the larger tool landscape.

Designing a search interface for a Spanish learner spoken corpus: the end-user's evaluation

Leonardo Campillos Llanos

This article summarizes the evaluation process of an interface under development to consult an oral corpus of learners of Spanish as a Foreign Language. The databank comprises 40 interviews with students with over 9 different mother tongues collected for Error Analysis. XML mark-up is used to code the information about the learners and their errors (with an explanation), and the search tool makes it is possible to look up these errors and to listen to the utterances where they appear. The formative evaluation was performed to improve the interface during the design stage by means of a questionnaire which addressed issues related to the teachers' beliefs about languages, their opinion about the Error Analysis methodology, and specific points about the interface design and usability. The results unveiled some deficiencies of the current prototype as well as the interests of the teaching professionals which should be considered to bridge the gap between technology development and its pedagogical applications.

P3 - Lexical Resources

Wednesday, May 23, 11:35

Chairperson: **Adam Kilgarriff** Poster Session

Dictionary Look-up with Katakana Variant Recognition

Satoshi Sato

The Japanese language has rich variety and quantity of word variant. Since 1980s, it has been recognized that this richness becomes an obstacle against natural language processing. A complete solution, however, has not been presented yet. This paper proposes a method to recognize Katakana variants–a major type of word variant in Japanese–in the process of dictionary look-up. For a given set of variant generation rules, the method executes variant generation and entry retrieval simultaneously and efficiently. We have developed the seven-layered rule set (216 rules in total) according to the specification manual of UniDic-2.1.0 and other sources. An experiment shows that the spelling-variant generator with 102 rules in the first five layers is almost perfect. Another experiment shows that the form-variant generator with all 216 rules is powerful and 77.7% of multiple spellings of Katakana loanwords are unnecessary (i.e., can be removed). This result means that the proposed method can drastically reduce the number of variants that we have to register into a dictionary in advance.

The Rocky Road towards a Swedish FrameNet - Creating SweFN

Karin Friberg Heppin and Maria Toporowska Gronostaj

The Swedish FrameNet project, SweFN, is a lexical resource under development, designed to support both humans and different applications within language technology, such as text generation, text understanding and information extraction. SweFN is constructed in line with the Berkeley FrameNet and the project is aiming to make it a free, full-scale, multi-functional lexical resource covering morphological, syntactic, and semantic descriptions of 50,000 entries. Frames populated by lexical units belonging to the general vocabulary dominate in SweFN, but there are also frames from the medical and the art domain. As Swedish is a language with very productive compounding, special attention is paid to semantic relations within the one word compounds which populate the frames. This is of relevance for understanding the meaning of the compounds and for capturing the semantic and syntactic alternations which are brought about in the course of compounding. SweFN is a component within a complex

of modern and historical lexicon resources named SweFN++, available at <http://spraakbanken.gu.se/eng/swefn>.

Capturing syntactico-semantic regularities among terms: An application of the FrameNet methodology to terminology

Marie-Claude L'Homme and Janine Pimentel

Terminological databases do not always provide detailed information on the linguistic behaviour of terms, although this is important for potential users such as translators or students. In this paper we describe a project that aims to fill this gap by proposing a method for annotating terms in sentences based on that developed within the FrameNet project (Ruppenhofer et al. 2010) and by implementing it in an online resource called DiCoInfo. We focus on the methodology we devised, and show with some preliminary results how similar actantial (i.e. argumental) structures can provide evidence for defining lexical relations in specific languages and capturing cross-linguistic equivalents. The paper argues that the syntactico-semantic annotation of the contexts in which the terms occur allows lexicographers to validate their intuitions concerning the linguistic behaviour of terms as well as interlinguistic relations between them. The syntactico-semantic annotation of contexts could, therefore, be considered a good starting point in terminology work that aims to describe the linguistic functioning of terms and offer a sounder basis to define interlinguistic relationships between terms that belong to different languages.

Developing LMF-XML Bilingual Dictionaries for Colloquial Arabic Dialects

David Graff and Mohamed Maamouri

The Linguistic Data Consortium and Georgetown University Press are collaborating to create updated editions of bilingual dictionaries that had originally been published in the 1960's for English-speaking learners of Moroccan, Syrian and Iraqi Arabic. In their first editions, these dictionaries used ad hoc Latin-alphabet orthography for each colloquial Arabic dialect, but adopted some proper- ties of Arabic-based writing (collation order of Arabic headwords, clitic attachment to word forms in example phrases); despite their common features, there are notable differences among the three books that impede comparisons across the dialects, as well as com- parisons of each dialect to Modern Standard Arabic. In updating these volumes, we use both Arabic script and International Pho- netic Alphabet orthographies; the former provides a common basis for word recognition across dialects, while the latter provides dialect-specific pronunciations. Our goal is to preserve the full content of the original publications, supplement the Arabic headword inventory with new usages, and produce a uniform lexicon structure expressible via the Lexical Markup Framework (LMF, ISO 24613). To this end, we developed a relational database schema that applies consistently to each dialect, and HTTP-based tools for searching, editing, workflow, review and inventory management.

UBY-LMF – A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF

Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek and Christian M. Meyer

We present UBY-LMF, an LMF-based model for large-scale, heterogeneous multilingual lexical-semantic resources (LSRs). UBY-LMF allows the standardization of LSRs down to a fine-grained level of lexical information by employing a large number of Data Categories from ISOCat. We evaluate UBY-LMF by converting nine LSRs in two languages to the corresponding format: the English WordNet, Wiktionary, Wikipedia, OmegaWiki, FrameNet and VerbNet and the German Wikipedia, Wiktionary and GermaNet. The resulting LSR, UBY (Gurevych et al., 2012), holds interoperable versions of all nine resources which can be queried by an easy to use public Java API. UBY-LMF covers a wide range of information types from expert-constructed and collaboratively constructed resources for English and German, also including links between different resources at the word sense level. It is designed to accommodate further resources and languages as well as automatically mined lexical-semantic knowledge.

Legal electronic dictionary for Czech

František Cvrček, Karel Pala and Pavel Rychlý

In the paper the results of the project of Czech Legal Electronic dictionary (PES) are presented. During the 4 year project the large legal terminological dictionary of Czech was created in the form of the electronic lexical database enriched with a hierarchical ontology of legal terms. It contains approx. 10,000 entries – legal terms together with their ontological relations and hypertext references. In the second part of the project the web interface based on the platform DEBII has been designed and implemented that allows users to browse and search effectively the database. At the same time the Czech Dictionary of Legal Terms will be generated from the database and later printed as a book. Inter-annotator's agreement in manual selection of legal terms was high – approx. 95 %.

Adaptive Dictionary for Bilingual Lexicon Extraction from Comparable Corpora

Amir Hazem and Emmanuel Morin

One of the main resources used for the task of bilingual lexicon extraction from comparable corpora is : the bilingual dictionary,

which is considered as a bridge between two languages. However, no particular attention has been given to this lexicon, except its coverage, and the fact that it can be issued from the general language, the specialised one, or a mix of both. In this paper, we want to highlight the idea that a better consideration of the bilingual dictionary by studying its entries and filtering the non-useful ones, leads to a better lexicon extraction and thus, reach a higher precision. The experiments are conducted on a medical domain corpora. The French-English specialised corpus 'breast cancer' of 1 million words. We show that the empirical results obtained with our filtering process improve the standard approach traditionally dedicated to this task and are promising for future work.

A New Twitter Verb Lexicon for Natural Language Processing

Jennifer Williams and Graham Katz

We describe in-progress work on the creation of a new lexical resource that contains a list of 486 verbs annotated with quantified temporal durations for the events that they describe. This resource is being compiled from more than 14 million tweets from the Twitter microblogging site. We are creating this lexicon of verbs and typical durations to address a gap in the available information that is represented in existing research. The data that is contained in this lexical resource is unlike any existing resources, which have been traditionally comprised from literature excerpts, news stories, and full-length weblogs. The kind of knowledge about how long an event lasts is crucial for natural language processing and is especially useful when the temporal duration of an event is implied. We are using data from Twitter because Twitter is a rich resource since people are publicly posting about real events and real durations of those events throughout the day.

P4 - Annotation and Corpora

Wednesday, May 23, 11:35

Chairperson: **Andreas Witt** Poster Session

Challenges in the development of annotated corpora of computer-mediated communication in Indian Languages: A Case of Hindi

Ritesh Kumar

The present paper describes an ongoing effort to compile and annotate a large corpus of computer-mediated communication (CMC) in Hindi. It describes the process of the compilation of the corpus, the basic structure of the corpus and the annotation of the corpus and the challenges faced in the creation of such a corpus. It also gives a description of the technologies developed for the processing of the data, addition of the metadata and annotation of the corpus. Since it is a corpus of written communication, it provides quite a distinctive challenge for the annotation process. Besides POS annotation, it will also be annotated at higher levels of representation. Once completely developed it will be a very useful resource of Hindi for research in the areas of linguistics, NLP and other social sciences research related to communication, particularly computer-mediated communication..Besides this the challenges discussed here and the way they are tackled could be taken as the model for developing the corpus of computer-mediated communication in other Indian languages. Furthermore the technologies developed for the construction of this corpus will also be made available publicly.

Ontologies of Linguistic Annotation: Survey and perspectives

Christian Chiarcos

This paper announces the release of the Ontologies of Linguistic Annotation (OLiA). The OLiA ontologies represent a repository of annotation terminology for various linguistic phenomena on a great band-width of languages. This paper summarizes the results of five years of research, it describes recent developments and directions for further research.

A High-Quality Web Corpus of Czech

Johanka Spoustová and Miroslav Spousta

In our paper, we present main results of the Czech grant project Internet as a Language Corpus, whose aim was to build a corpus of Czech web texts and to develop and publicly release related software tools. Our corpus may not be the largest web corpus of Czech, but it maintains very good language quality due to high portion of human work involved in the corpus development process. We describe the corpus contents (2.65 billions of words divided into three parts – 450 millions of words from news and magazines articles, 1 billion of words from blogs, diaries and other non-reviewed literary units, 1.1 billion of words from discussions messages), particular steps of the corpus creation (crawling, HTML and boilerplate removal, near duplicates removal, language filtering) and its automatic language annotation (POS tagging, syntactic parsing). We also describe our software tools being released under an open source license, especially a fast linear-time module for removing near-duplicates on a paragraph level.

WebAnnotator, an Annotation Tool for Web Pages

Xavier Tannier

This article presents WebAnnotator, a new tool for annotating Web pages. WebAnnotator is implemented as a Firefox extension,

allowing annotation of both offline and inline pages. The HTML rendering fully preserved and all annotations consist in new HTML spans with specific styles. WebAnnotator provides an easy and general-purpose framework and is made available under CeCILL free license (close to GNU GPL), so that use and further contributions are made simple. All parts of an HTML document can be annotated: text, images, videos, tables, menus, etc. The annotations are created by simply selecting a part of the document and clicking on the relevant type and subtypes. The annotated elements are then highlighted in a specific color. Annotation schemas can be defined by the user by creating a simple DTD representing the types and subtypes that must be highlighted. Finally, annotations can be saved (HTML with highlighted parts of documents) or exported (in a machine-readable format).

Development of a Web-Scale Chinese Word N-gram Corpus with Parts of Speech Information

Chi-Hsin Yu, Yi-jie Tang and Hsin-Hsi Chen

Web provides a large-scale corpus for researchers to study the language usages in real world. Developing a web-scale corpus needs not only a lot of computation resources, but also great efforts to handle the large variations in the web texts, such as character encoding in processing Chinese web texts. In this paper, we aim to develop a web-scale Chinese word N-gram corpus with parts of speech information called NTU PN-Gram corpus using the ClueWeb09 dataset. We focus on the character encoding and some Chinese-specific issues. The statistics about the dataset is reported. We will make the resulting corpus a public available resource to boost the Chinese language processing.

CoALT: A SOFTWARE FOR COMPARING AUTOMATIC LABELLING TOOLS

Dominique Fohr and Odile Mella

Speech-text alignment tools are frequently used in speech technology and research. In this paper, we propose a GPL software CoALT (Comparing Automatic Labelling Tools) for comparing two automatic labellers or two speech-text alignment tools, ranking them and displaying statistics about their differences. The main feature of CoALT is that a user can define its own criteria for evaluating and comparing the speech-text alignment tools since the required quality for labelling depends on the targeted application. Beyond ranking, our tool provides useful statistics for each labeller and above all about their differences and can emphasize the drawbacks and advantages of each labeller. We have applied our software for the French and English languages but it can be used for another language by simply defining the list of the phonetic symbols and optionally a set of phonetic rules. In this paper we present the usage of the software for comparing two automatic labellers on the corpus TIMIT. Moreover, as automatic labelling tools are configurable (number of GMMs, phonetic lexicon, acoustic parameterisation), we then present how CoALT allows to determine the best parameters for our automatic labelling tool.

CAT: the CELCT Annotation Tool

Valentina Bartalesi Lenzi, Giovanni Moretti and Rachele Sprugnoli

This paper presents CAT - CELCT Annotation Tool, a new general-purpose web-based tool for text annotation developed by CELCT (Center for the Evaluation of Language and Communication Technologies). The aim of CAT is to make text annotation an intuitive, easy and fast process. In particular, CAT was created to support human annotators in performing linguistic and semantic text annotation and was designed to improve productivity and reduce time spent on this task. Manual text annotation is, in fact, a time-consuming activity, and conflicts may arise with the strict deadlines annotation projects are frequently subject to. Thanks to its adaptability and user-friendly interface, CAT can positively contribute to improve time management in annotation project. Further, the tool has a number of features which make it an easy-to-use tool for many types of annotations. Even if the first prototype of CAT has been used to perform temporal and event annotation following the It-TimeML specifications, the tool is general enough to be used for annotating a broad range of linguistic and semantic phenomena. CAT is freely available for research purposes.

ROMBAC: The Romanian Balanced Annotated Corpus

Radu Ion, Elena Irimia, Dan Ştefănescu and Dan Tufiş

This article describes the collecting, processing and validation of a large balanced corpus for Romanian. The annotation types and structure of the corpus are briefly reviewed. It was constructed at the Research Institute for Artificial Intelligence of the Romanian Academy in the context of an international project (METANET4U). The processing covers tokenization, POS-tagging, lemmatization and chunking. The corpus is in XML format generated by our in-house annotation tools; the corpus encoding schema is XCES compliant and the metadata specification is conformant to the METANET recommendations. To the best of our knowledge, this is the first large and richly annotated corpus for Romanian. ROMBAC is intended to be the foundation of a linguistic environment containing a reference corpus for contemporary Romanian and a comprehensive collection of interoperable processing tools.

A French Fairy Tale Corpus syntactically and semantically annotated

Ismaïl El Maarouf and Jeanne Villaneau

Fairy tales, folktales and more generally children stories have lately attracted the Natural Language Processing (NLP) community. As such, very few corpora exist and linguistic resources are lacking. The work presented in this paper aims at filling this gap by presenting a syntactically and semantically annotated corpus. It focuses on the linguistic analysis of a Fairy Tales Corpus, and provides the description of the syntactic and semantic resources developed for Information Extraction. Resources include syntactic dependency relation annotation for 120 verbs; referential annotation, which is concerned with annotating each anaphoric occurrence and Proper Name with the most specific noun in the text; ontology matching for a substantial part of the nouns in the corpus; semantic role labelling for 41 verbs using the FrameNet database. The article also sums up previous analyses of this corpus and indicates possible uses of this corpus for the NLP community.

Iula2Standoff: a tool for creating standoff documents for the IULACT

Carlos Morell, Jorge Vivaldi and Núria Bel

Due to the increase in the number and depth of analyses required over the text, like entity recognition, POS tagging, syntactic analysis, etc. the annotation in-line has become unpractical. In Natural Language Processing (NLP) some emphasis has been placed in finding an annotation method to solve this problem. A possibility is the standoff annotation. With this annotation style it is possible to add new levels of annotation without disturbing exiting ones, with minimal knock on effects. This annotation will increase the possibility of adding more linguistic information as well as more possibilities for sharing textual resources. In this paper we present a tool developed in the framework of the European Metanet4u (Enhancing the European Linguistic Infrastructure, GA 270893) for creating a multi-layered XML annotation scheme, based on the GrAF proposal for standoff annotations.

ANALEC: a New Tool for the Dynamic Annotation of Textual Data

Frederic Landragin, Thierry Poibeau and Bernard Victorri

We introduce ANALEC, a tool which aim is to bring together corpus annotation, visualization and query management. Our main idea is to provide a unified and dynamic way of annotating textual data. ANALEC allows researchers to dynamically build their own annotation scheme and use the possibilities of scheme revision, data querying and graphical visualization during the annotation process. Each query result can be visualized using a graphical representation that puts forward a set of annotations that can be directly corrected or completed. Text annotation is then considered as a cyclic process. We show that statistics like frequencies and correlations make it possible to verify annotated data on the fly during the annotation. In this paper we introduce the annotation functionalities of ANALEC, some of the annotated data visualization functionalities, and three statistical modules: frequency, correlation and geometrical representations. Some examples dealing with reference and coreference annotation illustrate the main contributions of ANALEC.

The SYNC3 Collaborative Annotation Tool

Georgios Petasis

The huge amount of the available information in the Web creates the need of effective information extraction systems that are able to produce metadata that satisfy user's information needs. The development of such systems, in the majority of cases, depends on the availability of an appropriately annotated corpus in order to learn or evaluate extraction models. The production of such corpora can be significantly facilitated by annotation tools, that provide user-friendly facilities and enable annotators to annotate documents according to a predefined annotation schema. However, the construction of annotation tools that operate in a distributed environment is a challenging task: the majority of these tools are implemented as Web applications, having to cope with the capabilities offered by browsers. This paper describes the SYNC3 collaborative annotation tool, which implements an alternative architecture: it remains a desktop application, fully exploiting the advantages of desktop applications, but provides collaborative annotation through the use of a centralised server for storing both the documents and their metadata, and instance messaging protocols for communicating events among all annotators. The annotation tool is implemented as a component of the Ellogon language engineering platform, exploiting its extensive annotation engine, its cross-platform abilities and its linguistic processing components, if such a need arises. Finally, the SYNC3 annotation tool is distributed with an open source license, as part of the Ellogon platform.

Simplified guidelines for the creation of Large Scale Dialectal Arabic Annotations

Heba Elfardy and Mona Diab

The Arabic language is a collection of dialectal variants along with the standard form, Modern Standard Arabic (MSA). MSA is used in official Settings while the dialectal variants (DA) correspond

to the native tongue of the Arabic speakers. Arabic speakers typically code switch between DA and MSA, which is reflected extensively in written online social media. Automatic processing such Arabic genre is very difficult for automated NLP tools since the linguistic difference between MSA and DA is quite profound. However, no annotated resources exist for marking the regions of such switches in the utterance. In this paper, we present a simplified Set of guidelines for detecting code switching in Arabic on the word/token level. We use these guidelines in annotating a corpus that is rich in DA with frequent code switching to MSA. We present both a quantitative and qualitative analysis of the annotations.

O5 - Crowdsourcing (Special Session)

Wednesday, May 23, 14:45

Chairperson: **Karen Fort and Iryna Gurevych** Oral Session

Leveraging the Wisdom of the Crowds for the Acquisition of Multilingual Language Resources

Arno Scharl, Marta Sabou, Stefan Gindl, Walter Rafelsberger and Albert Weichselbraun

Games with a purpose are an increasingly popular mechanism for leveraging the wisdom of the crowds to address tasks which are trivial for humans but still not solvable by computer algorithms in a satisfying manner. As a novel mechanism for structuring human-computer interactions, a key challenge when creating them is motivating users to participate while generating useful and unbiased results. This paper focuses on important design choices and success factors of effective games with a purpose. Our findings are based on lessons learned while developing and deploying Sentiment Quiz, a crowdsourcing application for creating sentiment lexicons (an essential component of most sentiment detection algorithms). We describe the goals and structure of the game, the underlying application framework, the sentiment lexicons gathered through crowdsourcing, as well as a novel approach to automatically extend the lexicons by means of a bootstrapping process. Such an automated extension further increases the efficiency of the acquisition process by limiting the number of terms that need to be gathered from the game participants.

Experiences in Resource Generation for Machine Translation through Crowdsourcing

Anoop Kunchukuttan, Shourya Roy, Pratik Patel, Kushal Ladha, Somya Gupta, Mitesh M. Khapra and Pushpak Bhattacharyya

The logistics of collecting resources for Machine Translation (MT) has always been a cause of concern for some of the resource deprived languages of the world. The recent advent of crowdsourcing platforms provides an opportunity to explore the large scale generation of resources for MT. However, before venturing into this mode of resource collection, it is important to understand the various factors such as, task design, crowd motivation, quality control, etc. which can influence the success of such a crowd sourcing venture. In this paper, we present our experiences based on a series of experiments performed. This is an attempt to provide a holistic view of the different facets of translation crowd sourcing and identifying key challenges which need to be addressed for building a practical crowdsourcing solution for MT.

Irony and Sarcasm: Corpus Generation and Analysis Using Crowdsourcing

Elena Filatova

The ability to reliably identify sarcasm and irony in text can improve the performance of many Natural Language Processing (NLP) systems including summarization, sentiment analysis, etc. The existing sarcasm detection systems have focused on identifying sarcasm on a sentence level or for a specific phrase. However, often it is impossible to identify a sentence containing sarcasm without knowing the context. In this paper we describe a corpus generation experiment where we collect regular and sarcastic Amazon product reviews. We perform qualitative and quantitative analysis of the corpus. The resulting corpus can be used for identifying sarcasm on two levels: a document and a text utterance (where a text utterance can be as short as a sentence and as long as a whole document).

Supervised Topical Key Phrase Extraction of News Stories using Crowdsourcing, Light Filtering and Co-reference Normalization

Luís Marujo, Anatole Gershman, Jaime Carbonell, Robert Frederking and João P. Neto

Fast and effective automated indexing is critical for search and personalized services. Key phrases that consist of one or more words and represent the main concepts of the document are often used for the purpose of indexing. In this paper, we investigate the use of additional semantic features and pre-processing steps to improve automatic key phrase extraction. These features include the use of signal words and freebase categories. Some of these features lead to significant improvements in the accuracy of the results. We also experimented with 2 forms of document pre-processing that we call light filtering and co-reference normalization. Light filtering removes sentences from the document, which are judged peripheral to its main content. Co-reference normalization unifies several written forms of the same

named entity into a unique form. We also needed a "Gold Standard" – a set of labeled documents for training and evaluation. While the subjective nature of key phrase selection precludes a true "Gold Standard", we used Amazon's Mechanical Turk service to obtain a useful approximation. Our data indicates that the biggest improvements in performance were due to shallow semantic features, news categories, and rhetorical signals (nDCG 78.47% vs. 68.93%). The inclusion of deeper semantic features such as Freebase sub-categories was not beneficial by itself, but in combination with pre-processing, did cause slight improvements in the nDCG scores.

O6 - Dialogue and Multimodality

Wednesday, May 23, 14:45

Chairperson: **Jimmy Kunzmann** Oral Session

Constructive Interaction for Talking about Interesting Topics

Kristiina Jokinen and Graham Wilcock

The paper discusses mechanisms for topic management in conversations, concentrating on interactions where the interlocutors react to each other's presentation of new information and construct a shared context in which to exchange information about interesting topics. This is illustrated with a robot simulator that can talk about unrestricted (open-domain) topics that the human interlocutor shows interest in. Wikipedia is used as the source of information from which the robotic agent draws its world knowledge.

Using multimodal resources for explanation approaches in intelligent systems

Florian Nothdurft and Wolfgang Minker

In this work we show that there is a need of using multimodal resources during human-computer interaction (HCI) in intelligent systems. We propose that not only creating multimodal output for the user is important, but to take multimodal input resources into account for the decision when and how to interact. Especially the use of multimodal input resources for the decision when and how to provide assistance in HCI is important. The use of assistive functionalities like providing adaptive explanations to keep the user motivated and cooperative is more than a side-effect and demands a closer look. In this paper we introduce our approach on how to use multimodal input ressources in an adaptive and generic explanation pipeline. We do not only concentrate on using explanations as a way to manage user knowledge, but to maintain the cooperativeness, trust and motivation of the user to continue a healthy and well-structured HCI.

Multimodal Corpus of Multi-party Conversations in Second Language

Shota Yamasaki, Hirohisa Furukawa, Masafumi Nishida, Kristiina Jokinen and Seiichi Yamamoto

We developed a dialogue-based tutoring system for teaching English to Japanese students and plan to transfer the current software tutoring agent into an embodied robot in the hope that the robot will enrich conversation by allowing more natural interactions in small group learning situations. To enable smooth communication between an intelligent agent and the user, the agent must have realistic models on when to take turns, when to interrupt, and how to catch the partner's attention. For developing the realistic models applicable for computer assisted language learning systems, we also need to consider the differences between the mother tongue and second language that affect communication style. We collected a multimodal corpus of multi-party conversations in English as the second language to investigate the differences in communication styles. We describe our multimodal corpus and explore features of communication style e.g. filled pauses, and non-verbal information, such as eye-gaze, which show different characteristics between the mother tongue and second language.

The REX corpora: A collection of multimodal corpora of referring expressions in collaborative problem solving dialogues

Takenobu Tokunaga, Ryu Iida, Asuka Terai and Naoko Kuriyama

This paper describes a collection of multimodal corpora of referring expressions, the REX corpora. The corpora have two notable features, namely (1) they include time-aligned extra-linguistic information such as participant actions and eye-gaze on top of linguistic information, (2) dialogues were collected with various configurations in terms of the puzzle type, hinting and language. After describing how the corpora were constructed and sketching out each, we present an analysis of various statistics for the corpora with respect to the various configurations mentioned above. The analysis showed that the corpora have different characteristics in the number of utterances and referring expressions in a dialogue, the task completion time and the attributes used in the referring expressions. In this respect, we succeeded in constructing a collection of corpora that included a variety of characteristics by changing the configurations for each set of dialogues, as originally planned. The corpora are now under

preparation for publication, to be used for research on human reference behaviour.

ISO 24617-2: A semantically-based standard for dialogue annotation

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis and David Traum

This paper summarizes the latest, final version of ISO standard 24617-2 “Semantic annotation framework, Part 2: Dialogue acts". Compared to the preliminary version ISO DIS 24617-2:2010, described in Bunt et al. (2010), the final version additionally includes concepts for annotating rhetorical relations between dialogue units, defines a full-blown compositional semantics for the Dialogue Act Markup Language DiAML (resulting, as a side-effect, in a different treatment of functional dependence relations among dialogue acts and feedback dependence relations); and specifies an optimally transparent XML-based reference format for the representation of DiAML annotations, based on the systematic application of the notion of ‘ideal concrete syntax’. We describe these differences and briefly discuss the design and implementation of an incremental method for dialogue act recognition, which proves the usability of the ISO standard for automatic dialogue annotation.

O7 - Machine Translation and Language Resources (1)

Wednesday, May 23, 14:45

Chairperson: **Gregor Thurmair** Oral Session

Collecting and Using Comparable Corpora for Statistical Machine Translation

Inguna Skadina, Ahmet Aker, Nikos Mastropavlos, Fangzhong Su, Dan Tufiş, Mateja Verlic, Andrejs Vasiljevs, Bogdan Babych, Paul Clough, Robert Gaizauskas, Nikos Glaros, Monica Lestari Paramita and M ̄arcis Pinnis

Lack of sufficient parallel data for many languages and domains is currently one of the major obstacles to further advancement of automated translation. The ACCURAT project is addressing this issue by researching methods how to improve machine translation systems by using comparable corpora. In this paper we present tools and techniques developed in the ACCURAT project that allow additional data needed for statistical machine translation to be extracted from comparable corpora. We present methods and tools for acquisition of comparable corpora from the Web and other sources, for evaluation of the comparability of collected corpora, for multi-level alignment of comparable corpora and for extraction of lexical and terminological data for machine translation. Finally, we present initial evaluation results on the utility of collected corpora in domain-adapted machine translation and real-life applications.

Suffix Trees as Language Models

Casey Redd Kennington, Martin Kay and Annemarie Friedrich

Suffix trees are data structures that can be used to index a corpus. In this paper, we explore how some properties of suffix trees naturally provide the functionality of an n-gram language model with variable n. We explain these properties of suffix trees, which we leverage for our Suffix Tree Language Model (STLM) implementation and explain how a suffix tree implicitly contains the data needed for n-gram language modeling. We also discuss the kinds of smoothing techniques appropriate to such a model. We then show that our suffix-tree language model implementation is competitive when compared to the state-of-the-art language model SRILM (Stolke, 2002) in statistical machine translation experiments.

DGT-TM: A freely available Translation Memory in 22 languages

Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos and Patrick Schlüter

The European Commission’s (EC) Directorate General for Translation, together with the EC’s Joint Research Centre, is making available a large translation memory (TM; i.e. sentences and their professionally produced translations) covering twenty-two official European Union (EU) languages and their 231 language pairs. Such a resource is typically used by translation professionals in combination with TM software to improve speed and consistency of their translations. However, this resource has also many uses for translation studies and for language technology applications, including Statistical Machine Translation (SMT), terminology extraction, Named Entity Recognition (NER), multilingual classification and clustering, and many more. In this reference paper for DGT-TM, we introduce this new resource, provide statistics regarding its size, and explain how it was produced and how to use it.

Identifying Word Translations from Comparable Documents Without a Seed Lexicon

Reinhard Rapp, Serge Sharoff and Bogdan Babych

The extraction of dictionaries from parallel text corpora is an established technique. However, as parallel corpora are a scarce resource, in recent years the extraction of dictionaries using comparable corpora has obtained increasing attention. In order

to find a mapping between languages, almost all approaches suggested in the literature rely on a seed lexicon. The work described here achieves competitive results without requiring such a seed lexicon. Instead it presupposes mappings between comparable documents in different languages. For some common types of textual resources (e.g. encyclopedias or newspaper texts) such mappings are either readily available or can be established relatively easily. The current work is based on Wikipedias where the mappings between languages are determined by the authors of the articles. We describe a neural-network inspired algorithm which first characterizes each Wikipedia article by a number of keywords, and then considers the identification of word translations as a variant of word alignment in a noisy environment. We present results and evaluations for eight language pairs involving Germanic, Romanic, and Slavic languages as well as Chinese.

Large aligned treebanks for syntax-based machine translation

Gideon Kotzé, Vincent Vandeghinste, Scott Martens and Jörg Tiedemann

We present a collection of parallel treebanks that have been automatically aligned on both the terminal and the nonterminal constituent level for use in syntax-based machine translation. We describe how they were constructed and applied to a syntax- and example-based machine translation system called Parse and Corpus-Based Machine Translation (PaCo-MT). For the language pair Dutch to English, we present evaluation scores of both the nonterminal constituent alignments and the MT system itself, and in the latter case, compare them with those of Moses, a current state-of-the-art statistical MT system, when trained on the same data.

O8 - Corpus Processing and Infrastructure

Wednesday, May 23, 14:45

Chairperson: **Bente Maegaard** Oral Session

Korp – the corpus infrastructure of Språkbanken

Lars Borin, Markus Forsberg and Johan Roxendal

We present Korp, the corpus infrastructure of Språkbanken (the Swedish Language Bank). The infrastructure consists of three main components: the Korp corpus pipeline, the Korp backend, and the Korp frontend. The Korp corpus pipeline is used for importing corpora, annotating them, and then exporting the annotated corpora into different formats. An essential feature of the pipeline is the ability to leave existing annotations untouched, both structural and word level annotations, and to use the existing annotations as the foundation of other annotations. The Korp backend consists of a set of REST-based web services for searching in and retrieving information about the corpora. Finally, the Korp frontend is a graphical search interface that interacts with the Korp backend. The interface has been inspired by corpus search interfaces such as SketchEngine, Glossa, and DeepDict, and it uses State Chart XML (SCXML) in order to enable users to bookmark interaction states. We give a functional and technical overview of the three components, followed by a discussion of planned future work.

Annotation Trees: LDC's customizable, extensible, scalable, annotation infrastructure

Jonathan Wright, Kira Griffitt, Joe Ellis, Stephanie Strassel and Brendan Callahan

In recent months, LDC has developed a web-based annotation infrastructure centered around a tree model of annotations and a Ruby on Rails application called the LDC User Interface (LUI). The effort aims to centralize all annotation into this single platform, which means annotation is always available remotely, with no more software required than a web browser. While the design is monolithic in the sense of handling any number of annotation projects, it is also scalable, as it is distributed over many physical and virtual machines. Furthermore, minimizing customization was a core design principle, and new functionality can be plugged in without writing a full application. The creation and customization of GUIs is itself done through the web interface, without writing code, with the aim of eventually allowing project managers to create a new task without developer intervention. Many of the desirable features follow from the model of annotations as trees, and the operationalization of annotation as tree modification.

Building Large Corpora from the Web Using a New Efficient Tool Chain

Roland Schäfer and Felix Bildhauer

Over the last decade, methods of web corpus construction and the evaluation of web corpora have been actively researched. Prominently, the WaCky initiative has provided both theoretical results and a set of web corpora for selected European languages. We present a software toolkit for web corpus construction and a set of siginificantly larger corpora (up to over 9 billion tokens) built using this software. First, we discuss how the data should be collected to ensure that it is not biased towards certain hosts. Then, we describe our software toolkit which performs basic cleanups as well as boilerplate removal, simple connected text detection as well as shingling to remove duplicates from the corpora. We finally report evaluation results of the corpora built

so far, for example w.r.t. the amount of duplication contained and the text type/genre distribution. Where applicable, we compare our corpora to the WaCky corpora, since it is inappropriate, in our view, to compare web corpora to traditional or balanced corpora. While we use some methods applied by the WaCky initiative, we can show that we have introduced incremental improvements.

Annotated Bibliographical Reference Corpora in Digital Humanities

Young-Min Kim, Patrice Bellot, Elodie Faath and Marin Dacos

In this paper, we present new bibliographical reference corpora in digital humanities (DH) that have been developed under a research project, Robust and Language Independent Machine Learning Approaches for Automatic Annotation of Bibliographical References in DH Books supported by Google Digital Humanities Research Awards. The main target is the bibliographical references in the articles of Revues.org site, an oldest French online journal platform in DH field. Since the final object is to provide automatic links between related references and articles, the automatic recognition of reference fields like author and title is essential. These fields are therefore manually annotated using a set of carefully defined tags. After providing a full description of three corpora, which are separately constructed according to the difficulty level of annotation, we briefly introduce our experimental results on the first two corpora. A popular machine learning technique, Conditional Random Field (CRF) is used to build a model, which automatically annotates the fields of new references. In the experiments, we first establish a standard for defining features and labels adapted to our DH reference data. Then we show our new methodology against less structured references gives a meaningful result.

Building a 70 billion word corpus of English from ClueWeb

Jan Pomikálek, Miloš Jakubíček and Pavel Rychlý

This work describes the process of creation of a 70 billion word text corpus of English. We used an existing language resource, namely the ClueWeb09 dataset, as source for the corpus data. Processing such a vast amount of data presented several challenges, mainly associated with pre-processing (boilerplate cleaning, text de-duplication) and post-processing (indexing for efficient corpus querying using the CQL – Corpus Query Language) steps. In this paper we explain how we tackled them: we describe the tools used for boilerplate cleaning (jusText) and for de-duplication (onion) that was performed not only on full (document-level) duplicates but also on the level of near-duplicate texts. Moreover we show the impact of each of the performed pre-processing steps on the final corpus size. Furthermore we show how effective parallelization of the corpus indexation procedure was employed within the Manatee corpus management system and during computation of word sketches (one-page, automatic, corpus-derived summaries of a word's grammatical and collocational behaviour) from the resulting corpus.

P5 - Information Extraction (1)

Wednesday, May 23, 14:45

Chairperson: **Günter Neumann** Poster Session

A Gold Standard for Relation Extraction in the Food Domain

Michael Wiegand, Benjamin Roth, Eva Lasarcyk, Stephanie Köser and Dietrich Klakow

We present a gold standard for semantic relation extraction in the food domain for German. The relation types that we address are motivated by scenarios for which IT applications present a commercial potential, such as virtual customer advice in which a virtual agent assists a customer in a supermarket in finding those products that satisfy their needs best. Moreover, we focus on those relation types that can be extracted from natural language text corpora, ideally content from the internet, such as web forums, that are easy to retrieve. A typical relation type that meets these requirements are pairs of food items that are usually consumed together. Such a relation type could be used by a virtual agent to suggest additional products available in a shop that would potentially complement the items a customer has already in their shopping cart. Our gold standard comprises structural data, i.e. relation tables, which encode relation instances. These tables are vital in order to evaluate natural language processing systems that extract those relations.

Textual Characteristics for Language Engineering

Mathias Bank, Robert Remus and Martin Schierle

Language statistics are widely used to characterize and better understand language. In parallel, the amount of text mining and information retrieval methods grew rapidly within the last decades, with many algorithms evaluated on standardized corpora, often drawn from newspapers. However, up to now there were almost no attempts to link the areas of natural language processing and language statistics in order to properly characterize those evaluation corpora, and to help others to pick the most appropriate algorithms for their particular corpus. We believe no results in the field of natural language processing should be published without quantitatively describing the used corpora. Only then the real value of proposed methods can be determined and

the transferability to corpora originating from different genres or domains can be estimated. We lay ground for a language engineering process by gathering and defining a set of textual characteristics we consider valuable with respect to building natural language processing systems. We carry out a case study for the analysis of automotive repair orders and explicitly call upon the scientific community to provide feedback and help to establish a good practice of corpus-aware evaluations.

Automatically Extracting Procedural Knowledge from Instructional Texts using Natural Language Processing

Ziqi Zhang, Philip Webster, Victoria Uren, Andrea Varga and Fabio Ciravegna

Procedural knowledge is the knowledge required to perform certain tasks, and forms an important part of expertise. A major source of procedural knowledge is natural language instructions. While these readable instructions have been useful learning resources for human, they are not interpretable by machines. Automatically acquiring procedural knowledge in machine interpretable formats from instructions has become an increasingly popular research topic due to their potential applications in process automation. However, it has been insufficiently addressed. This paper presents an approach and an implemented system to assist users to automatically acquire procedural knowledge in structured forms from instructions. We introduce a generic semantic representation of procedures for analysing instructions, using which natural language techniques are applied to automatically extract structured procedures from instructions. The method is evaluated in three domains to justify the generality of the proposed semantic representation as well as the effectiveness of the implemented automatic system.

Evolution of Event Designation in Media: Preliminary Study

Xavier Tannier, Véronique Moriceau, Béatrice Arnulphy and Ruixin He

Within the general purpose of information extraction, detection of event descriptions is often an important clue. An important characteristic of event designation in texts, and especially in media, is that it changes over time. Understanding how these designations evolve is important in information retrieval and information extraction. Our first hypothesis is that, when an event first occurs, media relate it in a very descriptive way (using verbal designations) whereas after some time, they use shorter nominal designations instead. Our second hypothesis is that the number of different nominal designations for an event tends to stabilize itself over time. In this article, we present our methodology concerning the study of the evolution of event designations in French documents from the news agency AFP. For this preliminary study, we focused on 7 topics which have been relatively important in France. Verbal and nominal designations of events have been manually annotated in manually selected topic-related passages. This French corpus contains a total of 2064 annotations. We then provide preliminary interesting statistical results and observations concerning these evolutions.

CLTC: A Chinese-English Cross-lingual Topic Corpus

Yunqing Xia, Guoyu Tang, Peng Jin and Xia Yang

Cross-lingual topic detection within text is a feasible solution to resolving the language barrier in accessing the information. This paper presents a Chinese-English cross-lingual topic corpus (CLTC), in which 90,000 Chinese articles and 90,000 English articles are organized within 150 topics. Compared with TDT corpora, CLTC has three advantages. First, CLTC is bigger in size. This makes it possible to evaluate the large-scale cross-lingual text clustering methods. Second, articles are evenly distributed within the topics. Thus it can be used to produce test datasets for different purposes. Third, CLTC can be used as a cross-lingual comparable corpus to develop methods for cross-lingual information access. A preliminary evaluation with CLTC corpus indicates that the corpus is effective in evaluating cross-lingual topic detection methods.

A Resource-light Approach to Phrase Extraction for English and German Documents from the Patent Domain and User Generated Content

Julia Maria Schulz, Daniela Becks, Christa Womser-Hacker and Thomas Mandl

In order to extract meaningful phrases from corpora (e. g. in an information retrieval context) intensive knowledge of the domain in question and the respective documents is generally needed. When moving to a new domain or language the underlying knowledge bases and models need to be adapted, which is often time-consuming and labor-intensive. This paper adresses the described challenge of phrase extraction from documents in different domains and languages and proposes an approach, which does not use comprehensive lexica and therefore can be easily transferred to new domains and languages. The effectiveness of the proposed approach is evaluated on user generated content and documents from the patent domain in English and German.

An Evaluation of the Effect of Automatic Preprocessing on Syntactic Parsing for Biomedical Relation Extraction

Md. Faisal Mahbub Chowdhury and Alberto Lavelli

Relation extraction (RE) is an important text mining task which is the basis for further complex and advanced tasks. In state-of-the-art RE approaches, syntactic information obtained through

parsing plays a crucial role. In the context of biomedical RE previous studies report usage of various automatic preprocessing techniques applied before parsing the input text. However, these studies do not specify to what extent such techniques improve RE results and to what extent they are corpus specific as well as parser specific. In this paper, we aim at addressing these issues by using various preprocessing techniques, two syntactic tree kernel based RE approaches and two different parsers on 5 widely used benchmark biomedical corpora of the protein-protein interaction (PPI) extraction task. We also provide analyses of various corpus characteristics to verify whether there are correlations between these characteristics and the RE results obtained. These analyses of corpus characteristics can be exploited to compare the 5 PPI corpora.

Evaluation of Unsupervised Information Extraction

Wei Wang, Romaric Besançon, Olivier Ferret and Brigitte Grau

Unsupervised methods gain more and more attention nowadays in information extraction area, which allows to design more open extraction systems. In the domain of unsupervised information extraction, clustering methods are of particular importance. However, evaluating the results of clustering remains difficult at a large scale, especially in the absence of reliable reference. On the basis of our experiments on unsupervised relation extraction, we first discuss in this article how to evaluate clustering quality without a reference by relying on internal measures. Then we propose a method, supported by a dedicated annotation tool, for building a set of reference clusters of relations from a corpus. Moreover, we apply it to our experimental framework and illustrate in this way how to build a significant reference for unsupervised relation extraction, more precisely made of 80 clusters gathering more than 4,000 relation instances, in a short time. Finally, we present how such reference is exploited for the evaluation of clustering with external measures and analyze the results of the application of these measures to the clusters of relations produced by our unsupervised relation extraction system.

Extraction of unmarked quotations in Newspapers

Stéphanie Weiser and Patrick Watrin

This paper presents work in progress to automatically extract quotation sentences from newspaper articles. The focus is the extraction and annotation of unmarked quotation sentences. A linguistic study shows that unmarked quotation sentences can be formalised into 16 patterns that can be used to develop an extraction grammar. The question of unmarked quotation boundaries identification is also raised as they are often ambiguous. An annotation scheme allowing to describe all the elements that can take place in a quotation sentence is defined. This paper presents the creation of two resources necessary to our system. A dictionary of verbs introducing quotations has been automatically built using a grammar of marked quotations sentences to identify the verbs able to introduce quotations. A grammar formalising the patterns of unmarked quotation sentences – using the tool Unitex, based on finite state machines – has been developed. A short experiment has been performed on two patterns and shows some promising results.

NgramQuery - Smart Information Extraction from Google N-gram using External Resources

Martin Aleksandrov and Carlo Strapparava

This paper describes the implementation of a generalized query language on Google Ngram database. This language allows for very expressive queries that exploit semantic similarity acquired both from corpora (e.g. LSA) and from WordNet, and phonetic similarity available from the CMU Pronouncing Dictionary. It contains a large number of new operators, which combined in a proper query can help users to extract n-grams having similarly close syntactic and semantic relational properties. We also characterize the operators with respect to their corpus affiliation and their functionality. The query syntax is considered next given in terms of Backus-Naur rules followed by a few interesting examples of how the tool can be used. We also describe the command-line arguments the user could input comparing them with the ones for retrieving n-grams through the interface of Google Ngram database. Finally we discuss possible improvements on the extraction process and some relevant query completeness issues.

P6 - Word Sense Disambiguation and Evaluation

Wednesday, May 23, 14:45

Chairperson: **Sanni Nimb** Poster Session

A voting scheme to detect semantic underspecification

Héctor Martínez Alonso, Núria Bel and Bolette Sandford Pedersen

The following work describes a voting system to automatically classify the sense selection of the complex types Location/Organization and Container/Content, which depend on regular polysemy, as described by the Generative Lexicon (Pustejovsky, 1995) . This kind of sense alternations

very often presents semantic underspecificacion between its two possible selected senses. This kind of underspecification is not traditionally contemplated in word sense disambiguation systems, as disambiguation systems are still coping with the need of a representation and recognition of underspecification (Pustejovsky, 2009) The data are characterized by the morphosyntactic and lexical enviroment of the headwords and provided as input for a classifier. The baseline decision tree classifier is compared against an eight-member voting scheme obtained from variants of the training data generated by modifications on the class representation and from two different classification algorithms, namely decision trees and k-nearest neighbors. The voting system improves the accuracy for the non-underspecified senses, but the underspecified sense remains difficult to identify

A Comparative Evaluation of Word Sense Disambiguation Algorithms for German

Verena Henrich and Erhard Hinrichs

The present paper explores a wide range of word sense disambiguation (WSD) algorithms for German. These WSD algorithms are based on a suite of semantic relatedness measures, including path-based, information-content-based, and gloss-based methods. Since the individual algorithms produce diverse results in terms of precision and thus complement each other well in terms of coverage, a set of combined algorithms is investigated and compared in performance to the individual algorithms. Among the single algorithms considered, a word overlap method derived from the Lesk algorithm that uses Wiktionary glosses and GermaNet lexical fields yields the best F-score of 56.36. This result is outperformed by a combined WSD algorithm that uses weighted majority voting and obtains an F-score of 63.59. The WSD experiments utilize the German wordnet GermaNet as a sense inventory as well as WebCAGe (short for: Web-Harvested Corpus Annotated with GermaNet Senses), a newly constructed, sense-annotated corpus for this language. The WSD experiments also confirm that WSD performance is lower for words with fine-grained sense distinctions compared to words with coarse-grained senses.

DutchSemCor: Targeting the ideal sense-tagged corpus

Piek Vossen, Attila Görög, Rubén Izquierdo and Antal Van den Bosch

Word Sense Disambiguation (WSD) systems require large sense-tagged corpora along with lexical databases to reach satisfactory results. The number of English language resources for developed WSD increased in the past years while most other languages are still under-resourced. The situation is no different for Dutch. In order to overcome this data bottleneck, the DutchSemCor project will deliver a Dutch corpus that is sense-tagged with senses from the Cornetto lexical database. In this paper, we discuss the different conflicting requirements for a sense-tagged corpus and our strategies to fulfill them. We report on a first series of experiments to sup- port our semi-automatic approach to build the corpus.

Mapping WordNet synsets to Wikipedia articles

Samuel Fernando and Mark Stevenson

Lexical knowledge bases (LKBs), such as WordNet, have been shown to be useful for a range of language processing tasks. Extending these resources is an expensive and time-consuming process. This paper describes an approach to address this problem by automatically generating a mapping from WordNet synsets to Wikipedia articles. A sample of synsets has been manually annotated with article matches for evaluation purposes. The automatic methods are shown to create mappings with precision of 87.8% and recall of 46.9%. These mappings can then be used as a basis for enriching WordNet with new relations based on Wikipedia links. The manual and automatically created data is available online.

A new semantically annotated corpus with syntactic-semantic and cross-lingual senses

Myriam Rakho, Éric Laporte and Matthieu Constant

In this article, we describe a new sense-tagged corpus for Word Sense Disambiguation. The corpus is constituted of instances of 20 French polysemous verbs. Each verb instance is annotated with three sense labels: (1) the actual translation of the verb in the english version of this instance in a parallel corpus, (2) an entry of the verb in a computational dictionary of French (the Lexicon-Grammar tables) and (3) a fine-grained sense label resulting from the concatenation of the translation and the Lexicon-Grammar entry.

Detection of Peculiar Word Sense by Distance Metric Learning with Labeled Examples

Minoru Sasaki and Hiroyuki Shinnou

For natural language processing on machines, resolving such peculiar usages would be particularly useful in constructing a dictionary and dataset for word sense disambiguation. Hence, it is necessary to develop a method to detect such peculiar examples of a target word from a corpus. Note that, hereinafter, we define a peculiar example as an instance in which the target word or phrase has a new meaning.In this paper, we proposed a new peculiar example detection method using distance metric learning from labeled example pairs. In this method, first, distance

metric learning is performed by large margin nearest neighbor classification for the training data, and new training data points are generated using the distance metric in the original space. Then, peculiar examples are extracted using the local outlier factor, which is a density-based outlier detection method, from the updated training and test data. The efficiency of the proposed method was evaluated on an artificial dataset and the Semeval-2010 Japanese WSD task dataset. The results showed that the proposed method has the highest number of properly detected instances and the highest F-measure value. This shows that the label information of training data is effective for density-based peculiar example detection. Moreover, an experiment on outlier detection using a classification method such as SVM showed that it is difficult to apply the classification method to outlier detection.

Using semi-experts to derive judgments on word sense alignment: a pilot study

Soojeong Eom, Markus Dickinson and Graham Katz

The overall goal of this project is to evaluate the performance of word sense alignment (WSA) systems, focusing on obtaining examples appropriate to language learners. Building a gold standard dataset based on human expert judgments is costly in time and labor, and thus we gauge the utility of using semi-experts in performing the annotation. In an online survey, we present a sense of a target word from one dictionary with senses from the other dictionary, asking for judgments of relatedness. We note the difficulty of agreement, yet the utility in using such results to evaluate WSA work. We find that one's treatment of related senses heavily impacts the results for WSA.

ATLIS: Identifying Locational Information in Text Automatically

John Vogel, Marc Verhagen and James Pustejovsky

ATLIS (short for " ATLIS Tags Locations in Strings") is a tool being developed using a maximum-entropy machine learning model for automatically identifying information relating to spatial and locational information in natural language text. It is being developed in parallel with the ISO-Space standard for annotation of spatial information (Pustejovsky, Moszkowicz & Verhagen 2011). The goal of ATLIS is to be able to take in a document as raw text and mark it up with ISO-Space annotation data, so that another program could use the information in a standardized format to reason about the semantics of the spatial information in the document. The tool (as well as ISO-Space itself) is still in the early stages of development. At present it implements a subset of the proposed ISO-Space annotation standard: it identifies expressions that refer to specific places, as well as identifying prepositional constructions that indicate a spatial relationship between two objects. In this paper, the structure of the ATLIS tool is presented, along with preliminary evaluations of its performance.

P7 - Multiword Expressions and Term Extraction

Wednesday, May 23, 14:45

Chairperson: **Karel Pala** Poster Session

Semi-Supervised Technical Term Tagging With Minimal User Feedback

Behrang QasemiZadeh, Paul Buitelaar, Tianqi Chen and Georgeta Bordea

In this paper, we address the problem of extracting technical terms automatically from an unannotated corpus. We introduce a technology term tagger that is based on Liblinear Support Vector Machines and employs linguistic features including Part of Speech tags and Dependency Structures, in addition to user feedback to perform the task of identification of technology related terms. Our experiments show the applicability of our approach as witnessed by acceptable results on precision and recall.

Linguistic knowledge for specialized text production

Miriam Buendía-Castro and Beatriz Sánchez-Cárdenas

This paper outlines a proposal for encoding and describing verb phrase constructions in the knowledge base on the environment EcoLexicon, with the objective of helping translators in specialized text production. In order to be able to propose our own template, the characteristics and limitations of the most representative terminographic resources that include phraseological information were analyzed, along with the theoretical background that underlies the verb meaning argument structure in EcoLexicon. Our description provides evidence of the fact that this kind of entry structure can be easily encoded in other languages.

In the same boat and other idiomatic seafaring expressions

Rita Marinelli and Laura Cignoni

This paper reports on a research carried out at the Institute for Computational Linguistics (ILC) on a set of idiomatic nautical expressions in Italian and English. A total of 200 Italian expressions were first selected and examined, using both monolingual and bilingual dictionaries, as well as specific lexicographical works dealing with the subject of idiomaticity,

especially of the maritime type, and a similar undertaking was then conducted for the English expressions. We discuss the possibility of including both the Italian and English idiomatic expressions in the semantic database Mariterm, which contains terms belonging to the maritime domain. We describe the terminological database and the way in which the idiomatic expressions can be organised within the system, so that, similarly to the other synsets, they are connected to other concepts represented in the database, but at the same time continue to belong to a group of particular linguistic expressions. Furthermore, we study similarities and differences in meaning and usage of some idiomatic expressions in the two languages.

Association Norms of German Noun Compounds

Sabine Schulte im Walde, Susanne Borgwaldt and Ronny Jauch

This paper introduces association norms of German noun compounds as a lexical semantic resource for cognitive and computational linguistics research on compositionality. Based on an existing database of German noun compounds, we collected human associations to the compounds and their constituents within a web experiment. The current study describes the collection process and a part-of-speech analysis of the association resource. In addition, we demonstrate that the associations provide insight into the semantic properties of the compounds, and perform a case study that predicts the degree of compositionality of the experiment compound nouns, as relying on the norms. Applying a comparatively simple measure of association overlap, we reach a Spearman rank correlation coefficient of rs=0.5228; p<000001, when comparing our predictions with human judgements.

Medical Term Extraction in an Arabic Medical Corpus

Doaa Samy, Antonio Moreno-Sandoval, Conchi Bueno-Díaz, Marta Garrote-Salazar and José M. Guirao

This paper tests two different strategies for medical term extraction in an Arabic Medical Corpus. The experiments and the corpus are developed within the framework of Multimedica project funded by the Spanish Ministry of Science and Innovation and aiming at developing multilingual resources and tools for processing of newswire texts in the Health domain. The first experiment uses a fixed list of medical terms, the second experiment uses a list of Arabic equivalents of very limited list of common Latin prefix and suffix used in medical terms. Results show that using equivalents of Latin suffix and prefix outperforms the fixed list. The paper starts with an introduction, followed by a description of the state-of-art in the field of Arabic Medical Language Resources (LRs). The third section describes the corpus and its characteristics. The fourth and the fifth sections explain the lists used and the results of the experiments carried out on a sub-corpus for evaluation. The last section analyzes the results outlining the conclusions and future work.

Evaluating the Impact of External Lexical Resources into a CRF-based Multiword Segmenter and Part-of-Speech Tagger

Matthieu Constant and Isabelle Tellier

This paper evaluates the impact of external lexical resources into a CRF-based joint Multiword Segmenter and Part-of-Speech Tagger. We especially show different ways of integrating lexicon-based features in the tagging model. We display an absolute gain of 0.5% in terms of f-measure. Moreover, we show that the integration of lexicon-based features significantly compensates the use of a small training corpus.

Adapting and evaluating a generic term extraction tool

Anita Gojun, Ulrich Heid, Bernd Weißbach, Carola Loth and Insa Mingers

We present techniques for monolingual term candidate extraction which are being developed in the EU project TTC. We designed an application for German and English data that serves as a first evaluation of the methods for terminology extraction used in the project. The application situation highlighted the need for tools to handle lemmatization errors and to remove incomplete word sequences from multi-word term candidate lists, as well as the fact that the provision of German citation forms requires more morphological knowledge than TTC's slim approach can provide. We show a detailed evaluation of our extraction results and discuss the method for the evaluation of terminology extraction systems.

Evaluation of Classification Algorithms and Features for Collocation Extraction in Croatian

Mladen Karan, Jan Šnajder and Bojana Dalbelo Bašić

Collocations can be defined as words that occur together significantly more often than it would be expected by chance. Many natural language processing applications such as natural language generation, word sense disambiguation and machine translation can benefit from having access to information about collocated words. We approach collocation extraction as a classification problem where the task is to classify a given n-gram as either a collocation (positive) or a non-collocation (negative). Among the features used are word frequencies, classical association measures (Dice, PMI, chi2), and POS tags.

In addition, semantic word relatedness modeled by latent semantic analysis is also included. We apply wrapper feature subset selection to determine the best set of features. Performance of various classification algorithms is tested. Experiments are conducted on a manually annotated set of bigrams and trigrams sampled from a Croatian newspaper corpus. Best results obtained are 79.8 F1 measure for bigrams and 67.5 F1 measure for trigrams. The best classifier for bigrams was SVM, while for trigrams the decision tree gave the best performance. Features which contributed the most to overall performance were PMI, semantic relatedness, and POS information.

The Quaero Evaluation Initiative on Term Extraction

Thibault Mondary, Adeline Nazarenko, Haïfa Zargayouna and Sabine Barreaux

The Quaero program has organized a set of evaluations for terminology extraction systems in 2010 and 2011. Three objectives were targeted in this initiative: the first one was to evaluate the behavior and scalability of term extractors regarding the size of corpora, the second goal was to assess progress between different versions of the same systems, the last one was to measure the influence of corpus type. The protocol used during this initiative was a comparative analysis of 32 runs against a gold standard. Scores were computed using metrics that take into account gradual relevance. Systems produced by Quaero partners and publicly available systems were evaluated on pharmacology corpora composed of European Patents or abstracts of scientific articles, all in English. The gold standard was an unstructured version of the pharmacology thesaurus used by INIST-CNRS for indexing purposes. Most systems scaled with large corpora, contrasted differences were observed between different versions of the same systems and with better results on scientific articles than on patents. During the ongoing adjudication phase domain experts are enriching the thesaurus with terms found by several systems.

Using Noun Similarity to Adapt an Acceptability Measure for Persian Light Verb Constructions

Shiva Taslimipoor, Afsaneh Fazly and Ali Hamzeh

Light verb constructions (LVCs), such as take a walk and make a decision, are a common subclass of multiword expressions (MWEs), whose distinct syntactic and semantic properties call for a special treatment within a computational system. In particular, LVCs are formed semi-productively: often a semantically-general verb (such as take) combines with a number of semantically-similar nouns to form semantically-related LVCs, as in make a decision/choice/commitment. Nonetheless, there are restrictions as to which verbs combine with which class of nouns. A proper computational account of LVCs is even more important for languages such as Persian, in which most verbs are of the form of LVCs. Recently, there has been some work on the automatic identification of MWEs (including LVCs) in resource-rich languages, such as English and Dutch. We adapt such existing techniques for the automatic identification of LVCs in Persian, an under-resourced language. Specifically, we extend an existing statistical measure of the acceptability of English LVCs (Fazly et al., 2007) to make explicit use of semantic classes of noun, and show that such classes are in particular useful for determining the LVC acceptability of new combinations.

Identifying bilingual Multi-Word Expressions for Statistical Machine Translation

Dhouha Bouamor, Nasredine Semmar and Pierre Zweigenbaum

MultiWord Expressions (MWEs) repesent a key issue for numerous applications in Natural Language Processing (NLP) especially for Machine Translation (MT). In this paper, we describe a strategy for detecting translation pairs of MWEs in a French-English parallel corpus. In addition we introduce three methods aiming to integrate extracted bilingual MWE S in M OSES, a phrase based Statistical Machine Translation (SMT) system. We experimentally show that these textual units can improve translation quality.

Detecting Japanese Compound Functional Expressions using Canonical/Derivational Relation

Takafumi Suzuki, Yusuke Abe, Itsuki Toyota, Takehito Utsuro, Suguru Matsuyoshi and Masatoshi Tsuchiya

The Japanese language has various types of functional expressions. In order to organize Japanese functional expressions with various surface forms, a lexicon of Japanese functional expressions with hierarchical organization was compiled. This paper proposes how to design the framework of identifying more than 16,000 functional expressions in Japanese texts by utilizing hierarchical organization of the lexicon. In our framework, more than 16,000 functional expressions are roughly divided into canonical / derived functional expressions. Each derived functional expression is intended to be identified by referring to the most similar occurrence of its canonical expression. In our framework, contextual occurrence information of much fewer canonical expressions are expanded into the whole forms of derived expressions, to be utilized when identifying those derived expressions. We also empirically show that the proposed method can correctly identify more than 80% of the functional / content

usages only with less than 38,000 training instances of manually identified canonical expressions.

Building a database of French frozen adverbial phrases

Aude Grezka and Céline Poudat

The present paper gives an account of the approach we have led so far to build a database of frozen units. Although it has long been absent from linguistic studies and grammatical tradition, linguistic frozenness is currently a major research issue for linguistic studies, as frozen markers ensure the economy of the language system. The objective of our study is twofold: we first aim to build a comprehensive database of completely frozen units for the French language – what is traditionally called absolute or total frozenness. We started the project with the description of adverbial units – in the long term, we will also naturally describe adjectival, verbal and nominal phrases – and we will first present the database we have developed so far. This first objective is necessarily followed by the second one, which aims to assess the frozenness degree of the other units (i.e. relative frozenness). In this perspective, we resorted to two sets of methods: linguistic tests and statistical methods processed on two corpora (political and scientific discourse).

German Verb Patterns and Their Implementation in an Electronic Dictionary

Marc Luder

We describe an electronic lexical resource for German and the structure of its lexicon entries, notably the structure of verbal single-word and multi-word entries. The verb as the center of the sentence structure, as held by dependency models, is also a basic principle of the JAKOB narrative analysis application, for which the dictionary is the background. Different linguistic layers are combined for construing lexicon entries with a rich set of syntactic and semantic properties, suited to represent the syntactic and semantic behavior of verbal expressions (verb patterns), extracted from transcripts of real discourse, thereby lexicalizing the specific meaning of a specific verb pattern in a specific context. Verb patterns are built by the lexicographer by using a parser analyzing the input of a test clause and generating a machine-readable property string with syntactic characteristics and propositions for semantic characteristics grounded in an ontology. As an example, the German idiomatic expression“an den Karren fahren”(to come down hard on somebody) demonstrates the overall structure of a dictionary entry. The goal is to build unique dictionary entries (verb patterns) with reference to the whole of their properties.

P8 - Authoring Tools, Proofing

Wednesday, May 23, 14:45

Chairperson: **Catia Cucchiarini** Poster Session

Risk Analysis and Prevention: LELIE, a Tool dedicated to Procedure and Requirement Authoring

Flore Barcellini, Camille Albert, Corinne Grosse and Patrick Saint-Dizier

In this paper, we present the first phase of the LELIE project. A tool that detects business errors in technical documents such as procedures or requirements is introduced. The objective is to improve readability and to check for some elements of contents so that risks that could be entailed by misunderstandings or typos can be prevented. Based on a cognitive ergonomics analysis, we survey a number of frequently encountered types of errors and show how they can be detected using the <TextCoop> discourse analysis platform. We show how errors can be annotated, give figures on error frequencies and analyze how technical writers perceive our system.

A Framework for Spelling Correction in Persian Language Using Noisy Channel Model

Mohammad Hoseyn Sheykholeslam, Behrouz Minaei-Bidgoli and Hossein Juzi

There are several methods offered for spelling correction in Farsi (Persian) Language. Unfortunately no powerful framework has been implemented because of lack of a large training set in Farsi as an accurate model. A training set consisting of erroneous and related correction string pairs have been obtained from a large number of instances of the books each of which were typed two times in Computer Research Center of Islamic Sciences. We trained our error model using this huge set. In testing part after finding erroneous words in sample text, our program proposes some candidates for related correction. The paper focuses on describing the method of ranking related corrections. This method is customized version of Noisy Channel Spelling Correction for Farsi. This ranking method attempts to find intended correction c from a typo t, that maximizes P(c) P(t | c). In this paper different methods are described and analyzed to obtain a wide overview of the field. Our evaluation results show that Noisy Channel Model using our corpus and training set in this framework works more accurately and improves efficiently in comparison with other methods.

Conventional Orthography for Dialectal Arabic

Nizar Habash, Mona Diab and Owen Rambow

Dialectal Arabic (DA) refers to the day-to-day vernaculars spoken in the Arab world. DA lives side-by-side with the official language, Modern Standard Arabic (MSA). DA differs from MSA on all levels of linguistic representation, from phonology and morphology to lexicon and syntax. Unlike MSA, DA has no standard orthography since there are no Arabic dialect academies, nor is there a large edited body of dialectal literature that follows the same spelling standard. In this paper, we present CODA, a conventional orthography for dialectal Arabic; it is designed primarily for the purpose of developing computational models of Arabic dialects. We explain the design principles of CODA and provide a detailed description of its guidelines as applied to Egyptian Arabic.

Arabic Word Generation and Modelling for Spell Checking

Khaled Shaalan, Mohammed Attia, Pavel Pecina, Younes Samih and Josef van Genabith

Arabic is a language known for its rich and complex morphology. Although many research projects have focused on the problem of Arabic morphological analysis using different techniques and approaches, very few have addressed the issue of generation of fully inflected words for the purpose of text authoring. Available open-source spell checking resources for Arabic are too small and inadequate. Ayaspell, for example, the official resource used with OpenOffice applications, contains only 300,000 fully inflected words. We try to bridge this critical gap by creating an adequate, open-source and large-coverage word list for Arabic containing 9,000,000 fully inflected surface words. Furthermore, from a large list of valid forms and invalid forms we create a character-based tri-gram language model to approximate knowledge about permissible character clusters in Arabic, creating a novel method for detecting spelling errors. Testing of this language model gives a precision of 98.2% at a recall of 100%. We take our research a step further by creating a context-independent spelling correction tool using a finite-state automaton that measures the edit distance between input words and candidate corrections, the Noisy Channel Model, and knowledge-based rules. Our system performs significantly better than Hunspell in choosing the best solution, but it is still below the MS Spell Checker.

Similarity Ranking as Attribute for Machine Learning Approach to Authorship Identification

Jan Rygl and Aleš Horák

In the authorship identification task, examples of short writings of N authors and an anonymous document written by one of these N authors are given. The task is to determine the authorship of the anonymous text. Practically all approaches solved this problem with machine learning methods. The input attributes for the machine learning process are usually formed by stylistic or grammatical properties of individual documents or a defined similarity between a document and an author. In this paper, we present the results of an experiment to extend the machine learning attributes by ranking the similarity between a document and an author: we transform the similarity between an unknown document and one of the N authors to the order in which the author is the most similar to the document in the set of N authors. The comparison of similarity probability and similarity ranking was made using the Support Vector Machines algorithm. The results show that machine learning methods perform slightly better with attributes based on the ranking of similarity than with previously used similarity between an author and a document.

Spell Checking for Chinese

Shaohua Yang, Hai Zhao, Xiaolin Wang and Bao-liang Lu

This paper presents some novel results on Chinese spell checking. In this paper, a concise algorithm based on minimized-path segmentation is proposed to reduce the cost and suit the needs of current Chinese input systems. The proposed algorithm is actually derived from a simple assumption that spelling errors often make the number of segments larger. The experimental results are quite positive and implicitly verify the effectiveness of the proposed assumption. Finally, all approaches work together to output a result much better than the baseline with 12% performance improvement.

Spell Checking in Spanish: The Case of Diacritic Accents

Jordi Atserias, Maria Fuentes, Rogelio Nazar and Irene Renau

This article presents the problem of diacritic restoration (or diacritization) in the context of spell-checking, with the focus on an orthographically rich language such as Spanish. We argue that despite the large volume of work published on the topic of diacritization, currently available spell-checking tools have still not found a proper solution to the problem in those cases where both forms of a word are listed in the checker's dictionary. This is the case, for instance, when a word form exists with and without diacritics, such as continuo 'continuous' and continuó 'he/she/it continued', or when different diacritics make other word distinctions, as in continúo 'I continue'. We propose a very simple solution based on a word bigram model derived from correctly typed Spanish texts and evaluate the ability of this model to restore diacritics in artificial as well as real errors. The case of diacritics

is only meant to be an example of the possible applications for this idea, yet we believe that the same method could be applied to other kinds of orthographic or even grammatical errors. Moreover, given that no explicit linguistic knowledge is required, the proposed model can be used with other languages provided that a large normative corpus is available.

Incorporating an Error Corpus into a Spellchecker for Maltese

Michael Rosner, Albert Gatt, Andrew Attard and Jan Joachimsen

This paper discusses the ongoing development of a new Maltese spell checker, highlighting the methodologies which would best suit such a language. We thus discuss several previous attempts, highlighting what we believe to be their weakest point: a lack of attention to context. Two developments are of particular interest, both of which concern the availability of language resources relevant to spellchecking: (i) the Maltese Language Resource Server (MLRS) which now includes a representative corpus of c. 100M words extracted from diverse documents including the Maltese Legislation, press releases and extracts from Maltese web-pages and (ii) an extensive and detailed corpus of spelling errors that was collected whilst part of the MLRS texts were being prepared. We describe the structure of these resources as well as the experimental approaches focused on context that we are now in a position to adopt. We describe the framework within which a variety of different approaches to spellchecking and evaluation will be carried out, and briefly discuss the first baseline system we have implemented. We conclude the paper with a roadmap for future improvements.

O9 - Endangered Languages

Wednesday, May 23, 16:45

Chairperson: **Dafydd Gibbon** Oral Session

A Rule-based Morphological Analyzer for Murrinh-Patha

Melanie Seiss

Resource development mainly focuses on well-described languages with a large amount of speakers. However, smaller languages may also profit from language resources which can then be used in applications such as electronic dictionaries or computer-assisted language learning materials. The development of resources for such languages may face various challenges. Often, not enough data is available for a successful statistical approach and the methods developed for other languages may not be suitable for this specific language. This paper presents a morphological analyzer for Murrinh-Patha, a polysynthetic language spoken in the Northern Territory of Australia. While nouns in Murrinh-Patha only show minimal inflection, verbs in this language are very complex. The complexity makes it very difficult if not impossible to handle data in Murrinh-Patha with statistical, surface-oriented methods. I therefore present a rule-based morphological analyzer built in XFST and LEXC (Beesley and Karttunen, 2003) which can handle the inflection on nouns and adjectives as well as the complexities of the Murrinh-Patha verb.

Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages

Dirk Goldhahn, Thomas Eckart and Uwe Quasthoff

The Leipzig Corpora Collection offers free online access to 136 monolingual dictionaries enriched with statistical information. In this paper we describe current advances of the project in collecting and processing text data automatically for a large number of languages. Our main interest lies in languages of “low density”, where only few text data exists online. The aim of this approach is to create monolingual dictionaries and statistical information for a high number of new languages and to expand the existing dictionaries, opening up new possibilities for linguistic typology and other research. Focus of this paper will be set on the infrastructure for the automatic acquisition of large amounts of monolingual text in many languages from various sources. Preliminary results of the collection of text data will be presented. The mainly language-independent framework for preprocessing, cleaning and creating the corpora and computing the necessary statistics will also be depicted.

„Rendering Endangered Lexicons Interoperable through Standards Harmonization”: the RELISH project

Helen Aristar-Dry, Sebastian Drude, Menzo Windhouwer, Jost Gippert and Irina Nevskaya

The RELISH project promotes language-oriented research by addressing a two-pronged problem: (1) the lack of harmonization between digital standards for lexical information in Europe and America, and (2) the lack of interoperability among existing lexicons of endangered languages, in particular those created with the Shoebox/Toolbox lexicon building software. The cooperation partners in the RELISH project are the University of Frankfurt (FRA), the Max Planck Institute for Psycholinguistics (MPI Nijmegen), and Eastern Michigan University, the host of the Linguist List (ILIT). The project aims at harmonizing key European and American digital standards whose divergence

has hitherto impeded international collaboration on language technology for resource creation and analysis, as well as web services for archive access. Focusing on several lexicons of endangered languages, the project will establish a unified way of referencing lexicon structure and linguistic concepts, and develop a procedure for migrating these heterogeneous lexicons to a standards-compliant format. Once developed, the procedure will be generalizable to the large store of lexical resources involved in the LEGO and DoBeS projects.

Measuring the Divergence of Dependency Structures Cross-Linguistically to Improve Syntactic Projection Algorithms

Ryan Georgi, Fei Xia and William Lewis

Syntactic parses can provide valuable information for many NLP tasks, such as machine translation, semantic analysis, etc. However, most of the world's languages do not have large amounts of syntactically annotated corpora available for building parsers. Syntactic projection techniques attempt to address this issue by using parallel corpora between resource-poor and resource-rich languages, bootstrapping the resource-poor language with the syntactic analysis of the resource-rich language. In this paper, we investigate the possibility of using small, parallel, annotated corpora to automatically detect divergent structural patterns between two languages. These patterns can then be used to improve structural projection algorithms, allowing for better performing NLP tools for resource-poor languages, in particular those that may not have large amounts of annotated data necessary for traditional, fully-supervised methods. While this detection process is not exhaustive, we demonstrate that important instances of divergence are picked up with minimal prior knowledge of a given language pair.

O10 - Document Classification, Text Categorisation

Wednesday, May 23, 16:45

Chairperson: **Luca Dini** Oral Session

Measuring Interlanguage: Native Language Identification with L1-influence Metrics

Julian Brooke and Graeme Hirst

The task of native language (L1) identification suffers from a relative paucity of useful training corpora, and standard within-corpus evaluation is often problematic due to topic bias. In this paper, we introduce a method for L1 identification in second language (L2) texts that relies only on much more plentiful L1 data, rather than the L2 texts that are traditionally used for training. In particular, we do word-by-word translation of large L1 blog corpora to create a mapping to L2 forms that are a possible result of language transfer, and then use that information for unsupervised classification. We show this method is effective in several different learner corpora, with bigram features being particularly useful.

Distractorless Authorship Verification

John Noecker Jr and Michael Ryan

Authorship verification is the task of, given a document and a candi- date author, determining whether or not the document was written by the candi- date author. Traditional approaches to authorship verification have revolved around a "candidate author vs. everything else" approach. Thus, perhaps the most important aspect of performing authorship verification on a document is the development of an appropriate distractor set to represent "everything not the candidate author". The validity of the results of such experiments hinges on the ability to develop an appropriately representative set of distractor documents. Here, we propose a method for performing authorship verification without the use of a distractor set. Using only training data from the candidate author, we are able to perform authorship verification with high confidence (greater than 90% accuracy rates across a large corpus).

Correlation between Similarity Measures for Inter-Language Linked Wikipedia Articles

Monica Lestari Paramita, Paul Clough, Ahmet Aker and Robert Gaizauskas

Wikipedia articles in different languages have been mined to support various tasks, such as Cross-Language Information Retrieval (CLIR) and Statistical Machine Translation (SMT). Articles on the same topic in different languages are often connected by inter-language links, which can be used to identify similar or comparable content. In this work, we investigate the correlation between similarity measures utilising language-independent and language-dependent features and respective human judgments. A collection of 800 Wikipedia pairs from 8 different language pairs were collected and judged for similarity by two assessors. We report the development of this corpus and inter-assessor agreement between judges across the languages. Results show that similarity measured using language independent features is comparable to using an approach based on translating non-English documents. In both cases the correlation with human judgments is low but also dependent upon the language pair. The results and corpus generated from this work also provide insights into the measurement of cross-language similarity.

JRC Eurovoc Indexer JEX - A freely available multi-label categorisation tool

Ralf Steinberger, Mohamed Ebrahim and Marco Turchi

EuroVoc (2012) is a highly multilingual thesaurus consisting of over 6,700 hierarchically organised subject domains used by European Institutions and many authorities in Member States of the European Union (EU) for the classification and retrieval of official documents. JEX is JRC-developed multi-label classification software that learns from manually labelled data to automatically assign EuroVoc descriptors to new documents in a profile-based category-ranking task. The JEX release consists of trained classifiers for 22 official EU languages, of parallel training data in the same languages, of an interface that allows viewing and amending the assignment results, and of a module that allows users to re-train the tool on their own document collections. JEX allows advanced users to change the document representation so as to possibly improve the categorisation result through linguistic pre-processing. JEX can be used as a tool for interactive EuroVoc descriptor assignment to increase speed and consistency of the human categorisation process, or it can be used fully automatically. The output of JEX is a language-independent EuroVoc feature vector lending itself also as input to various other Language Technology tasks, including cross-lingual clustering and classification, cross-lingual plagiarism detection, sentence selection and ranking, and more.

O11 - Discourse (1)

Wednesday, May 23, 16:45

Chairperson: **Haïfa Zargayouna** Oral Session

Annotations for Power Relations on Email Threads

Vinodkumar Prabhakaran, Huzaifa Neralwala, Owen Rambow and Mona Diab

Social relations like power and influence are difficult concepts to define, but are easily recognizable when expressed. In this paper, we describe a multi-layer annotation scheme for social power relations that are recognizable from online written interactions. We introduce a typology of four types of power relations between dialog participants: hierarchical power, situational power, influence and control of communication. We also present a corpus of Enron emails comprising of 122 threaded conversations, manually annotated with instances of these power relations between participants. Our annotations also capture attempts at exercise of power or influence and whether those attempts were successful or not. In addition, we also capture utterance level annotations for overt display of power. We describe the annotation definitions using two example email threads from our corpus illustrating each type of power relation. We also present detailed instructions given to the annotators and provide various statistics on annotations in the corpus.

A Corpus for Research on Deliberation and Debate

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott and Joseph King

Deliberative, argumentative discourse is an important component of opinion formation, belief revision, and knowledge discovery; it is a cornerstone of modern civil society. Argumentation is productively studied in branches ranging from theoretical artificial intelligence to political rhetoric, but empirical analysis has suffered from a lack of freely available, unscripted argumentative dialogs. This paper presents the Internet Argument Corpus (IAC), a set of 390,704 posts in 11,800 discussions extracted from the online debate site 4forums.com. A 2866 thread/130,206 post extract of the corpus has been manually sided for topic of discussion, and subsets of this topic-labeled extract have been annotated for several dialogic and argumentative markers: degrees of agreement with a previous post, cordiality, audience-direction, combativeness, assertiveness, emotionality of argumentation, and sarcasm. As an application of this resource, the paper closes with a discussion of the relationship between discourse marker pragmatics, agreement, emotionality, and sarcasm in the IAC corpus.

Annotating Agreement and Disagreement in Threaded Discussion

Jacob Andreas, Sara Rosenthal and Kathleen McKeown

We introduce a new corpus of sentence-level agreement and disagreement annotations over LiveJournal and Wikipedia threads. This is the first agreement corpus to offer full-document annotations for threaded discussions. We provide a methodology for coding responses as well as an implemented tool with an interface that facilitates annotation of a specific response while viewing the full context of the thread. Both the results of an annotator questionnaire and high inter-annotator agreement statistics indicate that the annotations collected are of high quality.

Evaluation of Discourse Relation Annotation in the Hindi Discourse Relation Bank

Sudheer Kolachina, Rashmi Prasad, Dipti Misra Sharma and Aravind Joshi

We describe our experiments on evaluating recently proposed modifications to the discourse relation annotation scheme of the

Penn Discourse Treebank (PDTB), in the context of annotating discourse relations in Hindi Discourse Relation Bank (HDRB). While the proposed modifications were driven by the desire to introduce greater conceptual clarity in the PDTB scheme and to facilitate better annotation quality, our findings indicate that overall, some of the changes render the annotation task much more difficult for the annotators, as also reflected in lower inter-annotator agreement for the relevant sub-tasks. Our study emphasizes the importance of best practices in annotation task design and guidelines, given that a major goal of an annotation effort should be to achieve maximally high agreement between annotators. Based on our study, we suggest modifications to the current version of the HDRB, to be incorporated in our future annotation work.

O12 - Word Sense Disambiguation

Wednesday, May 23, 16:45

Chairperson: **Kiril Simov** Oral Session

Using Verb Subcategorization for Word Sense Disambiguation

Will Roberts and Valia Kordoni

We develop a model for predicting verb sense from subcategorization information and integrate it into SSI-Dijkstra, a wide-coverage knowledge-based WSD algorithm. Adding syntactic knowledge in this way should correct the current poor performance of WSD systems on verbs. This paper also presents, for the first time, an evaluation of SSI-Dijkstra on a standard data set which enables a comparison of this algorithm with other knowledge-based WSD systems. Our results show that our system is competitive with current graph-based WSD algorithms, and that the subcategorization model can be used to achieve better verb sense disambiguation performance.

Applying cross-lingual WSD to wordnet development

Marianna Apidianaki and Benoît Sagot

The automatic development of semantic resources constitutes an important challenge in the NLP community. The methods used generally exploit existing large-scale resources, such as Princeton WordNet, often combined with information extracted from multilingual resources and parallel corpora. In this paper we show how Cross-Lingual Word Sense Disambiguation can be applied to wordnet development. We apply the proposed method to WOLF, a free wordnet for French still under construction, in order to fill synsets that did not contain any literal yet and increase its coverage.

Discovering Missing Wikipedia Inter-language Links by means of Cross-lingual Word Sense Disambiguation

Els Lefever, Veronique Hoste and Martine De Cock

Wikipedia pages typically contain inter-language links to the corresponding pages in other languages. These links, however, are often incomplete. This paper describes a set of experiments in which the viability of discovering such missing inter-language links for ambiguous nouns by means of a cross-lingual Word Sense Disambiguation approach is investigated. The input for the inter-language link detection system is a set of Dutch pages for a given ambiguous noun and the output of the system is a set of links to the corresponding pages in three target languages (viz. French, Spanish and Italian). The experimental results show that although it is a very challenging task, the system succeeds to detect missing inter-language links between Wikipedia documents for a manually labeled test set. The final goal of the system is to provide a human editor with a list of possible missing links that should be manually verified.

Unsupervised Word Sense Disambiguation with Multilingual Representations

Erwin Fernandez-Ordonez, Rada Mihalcea and Samer Hassan

In this paper we investigate the role of multilingual features in improving word sense disambiguation. In particular, we explore the use of semantic clues derived from context translation to enrich the intended sense and therefore reduce ambiguity. Our experiments demonstrate up to 26% increase in disambiguation accuracy by utilizing multilingual features as compared to the monolingual baseline.

P9 - Morphology

Wednesday, May 23, 16:45

Chairperson: **Pushpak Bhattacharyya** Poster Session

First Steps towards the Semi-automatic Development of a Wordformation-based Lexicon of Latin

Marco Passarotti and Francesco Mambrini

Although lexicography of Latin has a long tradition dating back to ancient grammarians, and almost all Latin grammars devote to wordformation at least one part of the section(s) concerning morphology, none of the today available lexical

resources and NLP tools of Latin feature a wordformation-based organization of the Latin lexicon. In this paper, we describe the first steps towards the semi-automatic development of a wordformation-based lexicon of Latin, by detailing several problems occurring while building the lexicon and presenting our solutions. Developing a wordformation-based lexicon of Latin is nowadays of outmost importance, as the last years have seen a large growth of annotated corpora of Latin texts of different eras. While these corpora include lemmatization, morphological tagging and syntactic analysis, none of them features segmentation of the word forms and wordformation relations between the lexemes. This restricts the browsing and the exploitation of the annotated data for linguistic research and NLP tasks, such as information retrieval and heuristics in PoS tagging of unknown words.

PoliMorf: a (not so) new open morphological dictionary for Polish

Marcin Woliński, Marcin Miłkowski, Maciej Ogrodniczuk and Adam Przepiórkowski

This paper presents preliminary results of an effort aiming at the creation of a morphological dictionary of Polish, PoliMorf, available under a very liberal BSD-style license. The dictionary is a result of a merger of two existing resources, SGJP and Morfologik and was prepared within the CESAR/META-NET initiative. The work completed so far includes re-licensing of the two dictionaries and filling the new resource with the morphological data semi-automatically unified from both sources. The merging process is controlled by the collaborative dictionary development web application Kuźnia, also implemented within the project. The tool involves several advanced features such as using SGJP inflectional patterns for form generation, possibility of attaching dictionary labels and classification schemes to lexemes, dictionary source record and change tracking. Since SGJP and Morfologik are already used in a significant number of Natural Language Processing projects in Poland, we expect PoliMorf to become the Polish morphological dictionary of choice for many years to come.

Unsupervised acquisition of concatenative morphology

Lionel Nicolas, Jacques Farré and Cécile Darme

Among the linguistic resources formalizing a language, morphological rules are among those that can be achieved in a reasonable time. Nevertheless, since the construction of such resource can require linguistic expertise, morphological rules are still lacking for many languages. The automatized acquisition of morphology is thus an open topic of interest within the NLP field. We present an approach that allows to automatically compute, from raw corpora, a data-representative description of the concatenative mechanisms of a morphology. Our approach takes advantage of phenomena that are observable for all languages using morphological inflection and derivation but are more easy to exploit when dealing with concatenative mechanisms. Since it has been developed toward the objective of being used on as many languages as possible, applying this approach to a varied set of languages needs very few expert work. The results obtained for our first participation in the 2010 edition of MorphoChallenge have confirmed both the practical interest and the potential of the method.

Annotating and Learning Morphological Segmentation of Egyptian Colloquial Arabic

Emad Mohamed, Behrang Mohit and Kemal Oflazer

We present an annotation and morphological segmentation scheme for Egyptian Colloquial Arabic (ECA) in which we annotate user-generated content that significantly deviates from the orthographic and grammatical rules of Modern Standard Arabic and thus cannot be processed by the commonly used MSA tools. Using a per letter classification scheme in which each letter is classified as either a segment boundary or not, and using a memory-based classifier, with only word-internal context, prove effective and achieve a 92% exact match accuracy at the word level. The well-known MADA system achieves 81% while the per letter classification scheme using the ATB achieves 82%. Error analysis shows that the major problem is that of character ambiguity since the ECA orthography overloads the characters which would otherwise be more specific in MSA, like the differences between y () and Y () and A () , > ( ), and < () which are collapsed to y () and A () respectively or even totally confused and interchangeable. While normalization helps alleviate orthographic inconsistencies, it aggravates the problem of ambiguity.

First Results in a Study Evaluating Pre-annotation and Correction Propagation for Machine-Assisted Syriac Morphological Analysis

Paul Felt, Eric Ringger, Kevin Seppi, Kristian Heal, Robbie Haertel and Deryle Lonsdale

Manual annotation of large textual corpora can be cost-prohibitive, especially for rare and under-resourced languages. One potential solution is pre-annotation: asking human annotators to correct sentences that have already been annotated, usually by a machine. Another potential solution is correction propagation: using annotator corrections to bad pre-annotations to dynamically improve to the remaining pre-annotations within the current

sentence. The research presented in this paper employs a controlled user study to discover under what conditions these two machine-assisted annotation techniques are effective in increasing annotator speed and accuracy and thereby reducing the cost for the task of morphologically annotating texts written in classical Syriac. A preliminary analysis of the data indicates that pre-annotations improve annotator accuracy when they are at least 60% accurate, and annotator speed when they are at least 80% accurate. This research constitutes the first systematic evaluation of pre-annotation and correction propagation together in a controlled user study.

Evaluating Hebbian Self-Organizing Memories for Lexical Representation and Access

Claudia Marzi, Marcello Ferro, Claudia Caudai and Vito Pirrelli

The lexicon is the store of words in long-term memory. Any attempt at modelling lexical competence must take issues of string storage seriously. In the present contribution, we discuss a few desiderata that any biologically-inspired computational model of the mental lexicon has to meet, and detail a multi-task evaluation protocol for their assessment. The proposed protocol is applied to a novel computational architecture for lexical storage and acquisition, the"Topological Temporal Hebbian SOMs"(T2HSOMs), which are grids of topologically organised memory nodes with dedicated sensitivity to time-bound sequences of letters. These maps can provide a rigorous and testable conceptual framework within which to provide a comprehensive, multi-task protocol for testing the performance of Hebbian self-organising memories, and a comprehensive picture of the complex dynamics between lexical processing and the acquisition of morphological structure.

A Morphological Analyzer For Wolof Using Finite-State Techniques

Cheikh M. Bamba Dione

This paper reports on the design and implementation of a morphological analyzer for Wolof. The main motivation for this work is to obtain a linguistically motivated tool using finite-state techniques. The finite-state technology is especially attractive in dealing with human language morphologies. Finite-state transducers (FST) are fast, efficient and can be fully reversible, enabling users to perform analysis as well as generation. Hence, I use this approach to construct a new FST tool for Wolof, as a first step towards a computational grammar for the language in the Lexical Functional Grammar framework. This article focuses on the methods used to model complex morphological issues and on developing strategies to limit ambiguities. It discusses experimental evaluations conducted to assess the performance of the analyzer with respect to various statistical criteria. In particular, I also wanted to create morphosyntactically annotated resources for Wolof, obtained by automatically analyzing text corpora with a computational morphology.

IDENTIC Corpus: Morphologically Enriched Indonesian-English Parallel Corpus

Septina Dian Larasati

This paper describes the creation process of an Indonesian-English parallel corpus (IDENTIC). The corpus contains 45,000 sentences collected from different sources in different genres. Several manual text preprocessing tasks, such as alignment and spelling correction, are applied to the corpus to assure its quality. We also apply language specific text processing such as tokenization on both sides and clitic normalization on the Indonesian side. The corpus is available in two different formats: 'plain', stored in text format and 'morphologically enriched', stored in CoNLL format. Some parts of the corpus are publicly available at the IDENTIC homepage.

The Romanian Neuter Examined Through A Two-Gender N-Gram Classification System

Liviu P. Dinu, Vlad Niculae and Octavia-Maria Şulea

Romanian has been traditionally seen as bearing three lexical genders: masculine, feminine and neuter, although it has always been known to have only two agreement patterns (for masculine and feminine). A recent analysis of the Romanian gender system described in (Bateman and Polinsky, 2010), based on older observations, argues that there are two lexically unspecified noun classes in the singular and two different ones in the plural and that what is generally called neuter in Romanian shares the class in the singular with masculines, and the class in the plural with feminines based not only on agreement features but also on form. Previous machine learning classifiers that have attempted to discriminate Romanian nouns according to gender have so far taken as input only the singular form, presupposing the traditional tripartite analysis. We propose a classifier based on two parallel support vector machines using n-gram features from the singular and from the plural which outperforms previous classifiers in its high ability to distinguish the neuter. The performance of our system suggests that the two-gender analysis of Romanian, on which it is based, is on the right track.

UniDic for Early Middle Japanese: a Dictionary for Morphological Analysis of Classical Japanese

Toshinobu Ogiso, Mamoru Komachi, Yasuharu Den and Yuji Matsumoto

In order to construct an annotated diachronic corpus of Japanese, we propose to create a new dictionary for morphological

analysis of Early Middle Japanese (Classical Japanese) based on UniDic, a dictionary for Contemporary Japanese. Differences between the Early Middle Japanese and Contemporary Japanese, which prevent a naïve adaptation of UniDic to Early Middle Japanese, are found at the levels of lexicon, morphology, grammar, orthography and pronunciation. In order to overcome these problems, we extended dictionary entries and created a training corpus of Early Middle Japanese to adapt UniDic for Contemporary Japanese to Early Middle Japanese. Experimental results show that the proposed UniDic-EMJ, a new dictionary for Early Middle Japanese, achieves as high accuracy (97%) as needed for the linguistic research on lexicon and grammar in Japanese classical text analysis.

Recognition of Polish Derivational Relations Based on Supervised Learning Scheme

Maciej Piasecki, Radoslaw Ramocki and Marek Maziarz

The paper presents construction of *Derywator* – a language tool for the recognition of Polish derivational relations. It was built on the basis of machine learning in a way following the bootstrapping approach: a limited set of derivational pairs described manually by linguists in plWordNet is used to train *Derivator*. The tool is intended to be applied in semi-automated expansion of plWordNet with new instances of derivational relations. The training process is based on the construction of two transducers working in the opposite directions: one for prefixes and one for suffixes. Internal stem alternations are recognised, recorded in a form of mapping sequences and stored together with transducers. Raw results produced by *Derivator* undergo next corpus-based and morphological filtering. A set of derivational relations defined in plWordNet is presented. Results of tests for different derivational relations are discussed. A problem of the necessary corpus-based semantic filtering is analysed. The presented tool depends to a very little extent on the hand-crafted knowledge for a particular language, namely only a table of possible alternations and morphological filtering rules must be exchanged and it should not take longer than a couple of working days.

Reconstructing the Diachronic Morphology of Romanian from Dictionary Citations

Dan Cristea, Radu Simionescu and Gabriela Haja

This work represents a first step in the direction of reconstructing a diachronic morphology for Romanian. The main resource used in this task is the digital version of Romanian Language Dictionary (eDTLR). This resource offers various usage examples for its entries, citations extracted from popular Romanian texts, which often present diachronic and inflected forms of the word they are provided for. The concept of "word deformation" is introduced and classified into more categories. The research conducted aims at detecting one type of such deformations occurring in the citations – changes only in the stem of the current word, without the migration to another paradigm. An algorithm is presented which automatically infers old stem forms. This uses a paradigmatic data model of the current Romanian morphology. Having the inferred roots and the paradigms that they are part of, old flexion forms of the words can be deduced. Even more, by considering the years in which the citations were published, the inferred old word forms can be framed in certain periods of time, creating a great resource for research in the evolution of the Romanian language.

Generation of Verbal Stems in Derivationally Rich Language

Krešimir Šojat, Nives Mikelić Preradović and Marko Tadić

The paper presents a procedure for generating prefixed verbs in Croatian comprising combinations of one, two or three prefixes. The result of this generation process is a pool of derivationally valid prefixed verbs, although not necessarily occuring in corpora. The statistics of occurences of generated verbs in Croatian National Corpus has been calculated. Further usage of such language resource with generated potential verbs is also suggested, namely, enrichment of Croatian Morphological Lexicon, Croatian Wordnet and CROVALLEX.

A finite-state morphological transducer for Kyrgyz

Jonathan Washington, Mirlan Ipasov and Francis Tyers

This paper describes the development of a free/open-source finite-state morphological transducer for Kyrgyz. The transducer has been developed for morphological generation for use within a prototype Turkish ightarrowKyrgyz machine translation system, but has also been extensively tested for analysis. The finite-state toolkit used for the work was the Helsinki Finite-State Toolkit (HFST). The paper describes some issues in Kyrgyz morphology, the development of the tool, some linguistic issues encountered and how they were dealt with, and which issues are left to resolve. An evaluation is presented which shows that the transducer has medium-level coverage, between 82% and 87% on two freely available corpora of Kyrgyz, and high precision and recall over a manually verified test set.

AnIta: a powerful morphological analyser for Italian

Fabio Tamburini and Matias Melandri

In this paper we present AnIta, a powerful morphological analyser for Italian implemented within the framework of finite-state-automata models. It is provided by a large lexicon

containing more than 110,000 lemmas that enable it to cover relevant portions of Italian texts. We describe our design choices for the management of inflectional phenomena as well as some interesting new features to explicitly handle derivational and compositional processes in Italian, namely the wordform segmentation structure and Derivation Graph. Two different evaluation experiments, for testing coverage (Recall) and Precision, are described in detail, comparing the AnIta performances with some other freely available tools to handle Italian morphology. The experiments results show that the AnIta Morphological Analyser obtains the best performances among the tested systems, with Recall = 97.21% and Precision = 98.71%. This tool was a fundamental building block for designing a performant PoS-tagger and Lemmatiser for the Italian language that participated to two EVALITA evaluation campaigns ranking, in both cases, together with the best performing systems.

P10 - Prosody and Phonetics

Wednesday, May 23, 16:45

Chairperson: **Laurence Devillers** Poster Session

A topologic view of Topic and Focus marking in Italian

Gloria Gagliardi, Edoardo Lombardi Vallauri and Fabio Tamburini

Regularities in position and level of prosodic prominences associated to patterns of Information Structure are identified for some Italian varieties. The experiments' results suggest a possibly new structural hypothesis on the role and function of the main prominence in marking information patterns. (1) An abstract and merely structural,"topologic"concept of Prominence location can be conceived of, as endowed with the function of demarcation between units, before their culmination and"description". This may suffice to explain much of the process by which speakers interpret the IS of utterances in discourse. Further features, such as the specific intonational contours of the different IS units, may thus represent a certain amount of redundancy. (2) Real utterances do not always signal the distribution of Topic and Focus clearly. Acoustically, many remain underspecified in this respect. This is especially true for the distinction between Topic-Focus and Broad Focus, which indeed often has no serious effects on the progression of communicative dynamism in the subsequent discourse. (3) The consistency of such results with the law of least effort, and the very high percent of matching between perceptual evaluations and automatic measurement, seem to validate the used algorithm.

Comparison between two models of language for the automatic phonetic labeling of an undocumented language of the South-Asia: the case of Mo Piu

Geneviève Caelen-Haumont and Sethserey Sam

This paper aims at assessing the automatic labeling of an undocumented, unknown, unwritten and under-resourced language (Mo Piu) of the North Vietnam, by an expert phonetician. In the previous stage of the work, 7 sets of languages were chosen among Mandarin, Vietnamese, Khmer, English, French, to compete in order to select the best models of languages to be used for the phonetic labeling of Mo Piu isolated words. Two sets of languages (1degre Mandarin + French, 2degre Vietnamese + French) which got the best scores showed an additional distribution of their results. Our aim is now to study this distribution more precisely and more extensively, in order to statistically select the best models of languages and among them, the best sets of phonetic units which minimize the wrong phonetic automatic labeling.

MISTRAL+: A Melody Intonation Speaker Tonal Range semi-automatic Analysis using variable Levels

Benoît Weber, Geneviève Caelen-Haumont, Binh Hai Pham and Do-Dat Tran

This paper presents MISTRAL+, the upgraded version of an automatic tool created in 2004 named INTSMEL then MELISM. Since MELISM, the entire process has been modified in order to simplify and enhance the study of languages. MISTRAL+ is a combinaison of two modules: a Praat plugin MISTRAL_ Praat, and MISTRAL_ xls. For specific corpora, it performs phonological annotation based on the F0 variation in prominent words, but also in any chunk of speech, prominent or not. So this tool while being specialized can also be used as a generic one. Now among others, new functionalities allow to use API symbols while labeling, and to provide a semi-automatic melodic annotation in the frame of tonal languages. The program contains several functions which compute target points (or significant points) to model F0 contour, perform automatic annotation of different shapes and export all data in an xls file. In a first part of this paper, the MISTRAL+ functionalities will be described, and in a second part, an example of application will be presented about a study of the Mo Piu endangered language in the frame of the MICA Au Co Project.

Comparing performance of different set-covering strategies for linguistic content optimization in speech corpora

Nelly Barbot, Olivier Boeffard and Arnaud Delhay

Set covering algorithms are efficient tools for solving an optimal linguistic corpus reduction. The optimality of such a process is directly related to the descriptive features of the sentences of a reference corpus. This article suggests to verify experimentally the behaviour of three algorithms, a greedy approach and a lagrangian relaxation based one giving importance to rare events and a third one considering the Kullback-Liebler divergence between a reference and the ongoing distribution of events. The analysis of the content of the reduced corpora shows that the both first approaches stay the most effective to compress a corpus while guaranteeing a minimal content. The variant which minimises the Kullback-Liebler divergence guarantees a distribution of events close to a reference distribution as expected; however, the price for this solution is a much more important corpus. In the proposed experiments, we have also evaluated a mixed-approach considering a random complement to the smallest coverings.

Towards Fully Automatic Annotation of Audio Books for TTS

Olivier Boeffard, Laure Charonnat, Sébastien Le Maguer and Damien Lolive

Building speech corpora is a first and crucial step for every text-to-speech synthesis system. Nowadays, the use of statistical models implies the use of huge sized corpora that need to be recorded, transcribed, annotated and segmented to be usable. The variety of corpora necessary for recent applications (content, style, etc.) makes the use of existing digital audio resources very attractive. Among all available resources, audiobooks, considering their quality, are interesting. Considering this framework, we propose a complete acquisition, segmentation and annotation chain for audiobooks that tends to be fully automatic. The proposed process relies on a data structure, Roots, that establishes the relations between the different annotation levels represented as sequences of items. This methodology has been applied successfully on 11 hours of speech extracted from an audiobook. A manual check, on a part of the corpus, shows the efficiency of the process.

Statistical Evaluation of Pronunciation Encoding

Iris Merkus and Florian Schiel

In this study we investigate the idea to automatically evaluate newly created pronunciation encodings for being correct or containing a potential error. Using a cascaded triphone detector and phonotactical n-gram modeling with an optimal Bayesian threshold we classify unknown pronunciation transcripts into the classes 'probably faulty' or 'probably correct'. Transcripts tagged 'probably faulty' are forwarded to a manual inspection performed by an expert, while encodings tagged 'probably correct' are passed without further inspection. An evaluation of the new method on the German PHONOLEX lexical resource shows that with a tolerable error margin of approximately 3% faulty transcriptions a major reduction in work effort during the production of a new lexical resource can be achieved.

Annotating a corpus of human interaction with prosodic profiles – focusing on Mandarin repair/disfluency

Helen Kaiyun Chen

This study describes the construction of a manually annotated speech corpus that focuses on the sound profiles of repair/disfluency in Mandarin conversational interaction. Specifically, the paper focuses on how the tag set of prosodic profiles of the recycling repair culled from both audio-tapped and video-tapped, face-to-face Mandarin interaction are decided. By the methodology of both acoustic records and impressionistic judgements, 260 instances of Mandarin recycling repair are annotated with sound profiles including: pitch, duration, loudness, silence, and other observable prosodic cues (i.e. sound stretch and cut-offs). The study further introduces some possible applications of the current corpus, such as the implementation of the annotated data for analyzing the correlation between sound profiles of Mandarin repair and the interactional function of the repair. The goal of constructing the corpus is to facilitate an interdisciplinary study that concentrates on broadening the interactional linguistic theory by simultaneously paying close attention to the sound profiles emerged from interaction.

Prediction of Non-Linguistic Information of Spontaneous Speech from the Prosodic Annotation: Evaluation of the X-JToBI system

Kikuo Maekawa

Speakers' gender and age-group were predicted using the symbolic information of the X-JToBI prosodic labelling scheme as applied to the Core of the Corpus of Spontaneous Japanese (44 hours, 155 speakers, 201 talks). The correct prediction rate of speaker gender by means of logistic regression analysis was about 80%, and, the correct discrimination rate of speaker age-group (4 groups) by means of linear discriminant analysis was about 50 %. These results, in conjunction with the previously reported result of the prediction experiment of 4 speech registers from the X-JToBI information, shows convincingly the superiority of X-JToBI over the traditional J_ ToBI. Clarification of the mechanism

by which gender- and/or age-group information were reflected in the symbolic representations of prosody largely remains as open question, although some preliminary analyses were presented in the current paper.

Prosomarker: a prosodic analysis tool based on optimal pitch stylization and automatic syllabification

Antonio Origlia and Iolanda Alfano

Prosodic research in recent years has been supported by a number of automatic analysis tools aimed at simplifying the work that is requested to study intonation. The need to analyze large amounts of data and to inspect phenomena that are often ambiguous and difficult to model makes the prosodic research area an ideal application field for computer based processing. One of the main challenges in this field is to model the complex relations occurring between the segmental level, mainly in terms of syllable nuclei and boundaries, and the supra-segmental level, mainly in terms of tonal movements. The goal of our contribution is to provide a tool for automatic annotation of prosodic data, the Prosomarker, designed to give a visual representation of both segmental and suprasegmental events. The representation is intended to be as generic as possible to let researchers analyze specific phenomena without being limited by assumptions introduced by the annotation itself. A perceptual account of the pitch curve is provided along with an automatic segmentation of the speech signal into syllable-like segments and the tool can be used both for data exploration, in semi-automatic mode, and to process large sets of data, in automatic mode.

Designing French Tale Corpora for Entertaining Text To Speech Synthesis

David Doukhan, Sophie Rosset, Albert Rilliard, Christophe d'Alessandro and Martine Adda-Decker

Text and speech corpora for training a tale telling robot have been designed, recorded and annotated. The aim of these corpora is to study expressive storytelling behaviour, and to help in designing expressive prosodic and co-verbal variations for the artificial storyteller). A set of 89 children tales in French serves as a basis for this work. The tales annotation principles and scheme are described, together with the corpus description in terms of coverage and inter-annotator agreement. Automatic analysis of a new tale with the help of this corpus and machine learning is discussed. Metrics for evaluation of automatic annotation methods are discussed. A speech corpus of about 1 hour, with 12 tales has been recorded and aligned and annotated. This corpus is used for predicting expressive prosody in children tales, above the level of the sentence.

Open-Source Boundary-Annotated Corpus for Arabic Speech and Language Processing

Claire Brierley, Majdi Sawalha and Eric Atwell

A boundary-annotated and part-of-speech tagged corpus is a prerequisite for developing phrase break classifiers. Boundary annotations in English speech corpora are descriptive, delimiting intonation units perceived by the listener. We take a novel approach to phrase break prediction for Arabic, deriving our prosodic annotation scheme from Tajwīd (recitation) mark-up in the Qur'an which we then interpret as additional text-based data for computational analysis. This mark-up is prescriptive, and signifies a widely-used recitation style, and one of seven original styles of transmission. Here we report on version 1.0 of our Boundary-Annotated Qur'an dataset of 77430 words and 8230 sentences, where each word is tagged with prosodic and syntactic information at two coarse-grained levels. In (Sawalha et al., 2012), we use the dataset in phrase break prediction experiments. This research is part of a larger-scale project to produce annotation schemes, language resources, algorithms, and applications for Classical and Modern Standard Arabic.

A Phonemic Corpus of Polish Child-Directed Speech

Luc Boruta and Justyna Jastrzebska

Recent advances in modeling early language acquisition are due not only to the development of machine-learning techniques, but also to the increasing availability of data on child language and child-adult interaction. In the absence of recordings of child-directed speech, or when models explicitly require such a representation for training data, phonemic transcriptions are commonly used as input data. We present a novel (and to our knowledge, the first) phonemic corpus of Polish child-directed speech. It is derived from the Weist corpus of Polish, freely available from the seminal CHILDES database. For the sake of reproducibility, and to exemplify the typical trade-off between ecological validity and sample size, we report all preprocessing operations and transcription guidelines. Contributed linguistic resources include updated CHAT-formatted transcripts with phonemic transcriptions in a novel phonology tier, as well as by-product data, such as a phonemic lexicon of Polish. All resources are distributed under the LGPL-LR license.

P11 - Language Resource Infrastructures (1)

Wednesday, May 23, 16:45

Chairperson: **Daan Broeder** Poster Session

Smooth Sailing for STEVIN

Peter Spyns and Elisabeth D'Halleweyn

In this paper we report on the past evaluation of the STEVIN programme in the field of Human Language Technology for Dutch (HLTD). STEVIN was a 11.4 M euro programme that was jointly organised and financed by the Flemish and Dutch governments. The aim was to provide academia and industry with basic building blocks for a linguistic infrastructure for the Dutch language. An independent evaluation has been carried out. The evaluators concluded that the most important targets of the STEVIN programme have been achieved to a very high extent. In this paper, we summarise the context, the evaluation method, the resulting resources and the highlights of the STEVIN final evaluation.

Semantic metadata mapping in practice: the Virtual Language Observatory

Dieter van Uytvanck, Herman Stehouwer and Lari Lampen

In this paper we present the Virtual Language Observatory (VLO), a metadata-based portal for language resources. It is completely based on the Component Metadata (CMDI) and ISOcat standards. This approach allows for the use of heterogeneous metadata schemas while maintaining the semantic compatibility. We describe the metadata harvesting process, based on OAI-PMH, and the conversion from several formats (OLAC, IMDI and the CLARIN LRT inventory) to their CMDI counterpart profiles. Then we focus on some post-processing steps to polish the harvested records. Next, the ingestion of the CMDI files into the VLO facet browser is described. We also include an overview of the changes since the first version of the VLO, based on user feedback from the CLARIN community. Finally there is an overview of additional ideas and improvements for future versions of the VLO.

Aspects of a Legal Framework for Language Resource Management

Aditi Sharma Grover, Annamart Nieman, Gerhard Van Huyssteen and Justus Roux

The management of language resources requires several legal aspects to be taken into consideration. In this paper we discuss a number of these aspects which lead towards the formation of a legal framework for a language resources management agency. The legal framework entails examination of; the agency's stakeholders and the relationships that exist amongst them, the privacy and intellectual property rights that exist around the language resources offered by the agency, and the external (e.g. laws, acts, policies) and internal legal instruments (e.g. end user licence agreements) required for the agency's operation.

Introducing the Swedish Kelly-list, a new lexical e-resource for Swedish

Elena Volodina and Sofie Johansson Kokkinakis

Frequency lists and/or lexicons contain information about the words and their statistics. They tend to find their "readers" among language learners, language teachers, linguists and lexicographers. Making them available in electronic format helps to expand the target group to cover language engineers, computer programmers and other specialists working in such areas as information retrieval, spam filtering, text readability analysis, test generation etc. This article describes a new freely available electronic frequency list of modern Swedish which was created in the EU project KELLY. We provide a short description of the KELLY project; examine the methodological approach and mention some details on the compiling of the corpus from which the list has been derived. Further, we discuss the type of information the list contains; describe the steps for list generation; provide information on the coverage and some other statistics over the items in the list. Finally, some practical information on the license for the Swedish Kelly-list distribution is given; potential application areas are suggested; and future plans for its expansion are mentioned. We hope that with some publicity we can help this list find its users.

Texto4Science: a Quebec French Database of Annotated Short Text Messages

Philippe Langlais, Patrick Drouin, Amélie Paulus, Eugénie Rompré Brodeur and Florent Cottin

In October 2009, was launched the Quebec French part of the international sms4science project, called texto4science. Over a period of 10 months, we collected slightly more than 7000 SMSs that we carefully annotated. This database is now ready to be used by the community. The purpose of this article is to relate the efforts put into designing this database and provide some data analysis of the main linguistic phenomenon that we have annotated. We also report on a socio-linguistic survey we conducted within the project.

Recent Developments in CLARIN-NL

Jan Odijk

In this paper we describe recent developments in the CLARIN-NL project with the goal of sharing information on and

experiences in this project with the community outside of the Netherlands. We discuss a variety of subprojects to actually implement the infrastructure, to provide functionality for search in metadata and the actual data, resource curation and demonstration projects, the Data Curation Service, actions to improve semantic interoperability and coordinate work on it, involvement of CLARIN Data Providers, education and training, outreach activities, and cooperation with other projects. Based on these experiences, we provide some recommendations for related projects. The recommendations concern a variety of topics including the organisation of an infrastructure project as a function of the types of tasks that have to be carried out, involvement of the targeted users, metadata, semantic interoperability and the role of registries, measures to maximally ensure sustainability, and cooperation with similar projects in other countries.

A Metadata Editor to Support the Description of Linguistic Resources

Emanuel Dima, Christina Hoppermann, Erhard Hinrichs, Thorsten Trippel and Claus Zinn

Creating and maintaining metadata for various kinds of resources requires appropriate tools to assist the user. The paper presents the metadata editor ProFormA for the creation and editing of CMDI (Component Metadata Infrastructure) metadata in web forms. This editor supports a number of CMDI profiles currently being provided for different types of resources. Since the editor is based on XForms and server-side processing, users can create and modify CMDI files in their standard browser without the need for further processing. Large parts of ProFormA are implemented as web services in order to reuse them in other contexts and programs.

Fivehundredmillionandone Tokens. Loading the AAC Container with Text Resources for Text Studies.

Hanno Biber and Evelyn Breiteneder

The“AAC - Austrian Academy Corpus”is a diachronic German language digital text corpus of more than 500 million tokens. The text corpus has collected several thousands of texts representing a wide range of different text types. The primary research aim is to develop text language resources for the study of texts. For corpus linguistics and corpus based language research large text corpora need to be structured in a systematic way. For this structural purpose the AAC is making use of the notion of container. By container in the context of corpus research we understand a flexible system of pragmatic representation, manipulation, modification and structured storage of annotated items of text. The issue of representing a large corpus in formats that offer only limited space is paradigmatic for the general task of representing a language by just a small collection of text or a small sample of the language. Methods based upon structural normalization and standardization have to be developed in order to provide useful instruments for text studies.

The Common Orthographic Vocabulary of the Portuguese Language: a set of open lexical resources for a pluricentric language

José Pedro Ferreira, Maarten Janssen, Gladis Barcellos de Oliveira, Margarita Correia and Gilvan Müller de Oliveira

This paper outlines the design principles and choices, as well as the ongoing development process of the Common Orthographic Vocabulary of the Portuguese Language (VOC), a large scale electronic lexical database which was adopted by the Community of Portuguese-Speaking Countries' (CPLP) Instituto Internacional da Língua Portuguesa to implement a spelling reform that is currently taking place. Given the different available resources and lexicographic traditions within the CPLP countries, a range of different solutions was adopted for different countries and integrated into a common development framework. Although the publication of lexicographic resources to implement spelling reforms has always been done for Portuguese, VOC represents a paradigm change, switching from idiosyncratic, closed source, paper-format official resources to standardized, open, free, web-accessible and reusable ones. We start by outlining the context that justifies the resource development and its requirements, then focusing on the description of the methodology, workflow and tools used, showing how a collaborative project in a common web-based platform and administration interface make the creation of such a long-sought and ambitious project possible.

Creation of an Open Shared Language Resource Repository in the Nordic and Baltic Countries

Andrejs Vasiljevs, Markus Forsberg, Tatiana Gornostay, Dorte Haltrup Hansen, Kristín Jóhannsdóttir, Gunn Lyse, Krister Lindén, Lene Offersgaard, Sussi Olsen, Bolette Pedersen, Eiríkur Rögnvaldsson, Inguna Skadina, Koenraad De Smedt, Ville Oksanen and Roberts Rozis

The META-NORD project has contributed to an open infrastructure for language resources (data and tools) under the META-NET umbrella. This paper presents the key objectives of META-NORD and reports on the results achieved in the first year of the project. META-NORD has mapped and described the national language technology landscape in the Nordic and Baltic countries in terms of language use, language technology and resources, main actors in the academy, industry, government

and society; identified and collected the first batch of language resources in the Nordic and Baltic countries; documented, processed, linked, and upgraded the identified language resources to agreed standards and guidelines. The three horizontal multilingual actions in META-NORD are overviewed in this paper: linking and validating Nordic and Baltic wordnets, the harmonisation of multilingual Nordic and Baltic treebanks, and consolidating multilingual terminology resources across European countries. This paper also touches upon intellectual property rights for the sharing of language resources.

The LRE Map. Harmonising Community Descriptions of Resources

Nicoletta Calzolari, Riccardo Del Gratta, Gil Francopoulo, Joseph Mariani, Francesco Rubino, Irene Russo and Claudia Soria

Accurate and reliable documentation of Language Resources is an undisputable need: documentation is the gateway to discovery of Language Resources, a necessary step towards promoting the data economy. Language resources that are not documented virtually do not exist: for this reason every initiative able to collect and harmonise metadata about resources represents a valuable opportunity for the NLP community. In this paper we describe the LRE Map, reporting statistics on resources associated with LREC2012 papers and providing comparisons with LREC2010 data. The LRE Map, jointly launched by FLaReNet and ELRA in conjunction with the LREC 2010 Conference, is an instrument for enhancing availability of information about resources, either new or already existing ones. It wants to reinforce and facilitate the use of standards in the community. The LRE Map web interface provides the possibility of searching according to a fixed set of metadata and to view the details of extracted resources. The LRE Map is continuing to collect bottom-up input about resources from authors of other conferences through standard submission process. This will help broadening the notion of "language resources" and attract to the field neighboring disciplines that so far have been only marginally involved by the standard notion of language resources.

The META-SHARE Metadata Schema for the Description of Language Resources

Maria Gavrilidou, Penny Labropoulou, Elina Desipri, Stelios Piperidis, Haris Papageorgiou, Monica Monachini, Francesca Frontini, Thierry Declerck, Gil Francopoulo, Victoria Arranz and Valérie Mapelli

This paper presents a metadata model for the description of language resources proposed in the framework of the META-SHARE infrastructure, aiming to cover both datasets and tools/technologies used for their processing. It places the model in the overall framework of metadata models, describes the basic principles and features of the model, elaborates on the distinction between minimal and maximal versions thereof, briefly presents the integrated environment supporting the LRs description and search and retrieval processes and concludes with work to be done in the future for the improvement of the model.

Towards automation in using multi-modal language resources: compatibility and interoperability for multi-modal features in Kachako

Yoshinobu Kano

Use of language resources including annotated corpora and tools is not easy for users, as it requires expert knowledge to determine which resources are compatible and interoperable. Sometimes it requires programming skill in addition to the expert knowledge to make the resources compatible and interoperable when the resources are not created so. If a platform system could provide automation features for using language resources, users do not have to waste their time as the above issues are not necessarily essential for the users' goals. While our system, Kachako, provides such automation features for single-modal resources, multi-modal resources are more difficult to combine automatically. In this paper, we discuss designs of multi-modal resource compatibility and interoperability from such an automation point of view in order for the Kachako system to provide automation features of multi-modal resources. Our discussion is based on the UIMA framework, and focuses on resource metadata description optimized for ideal automation features while harmonizing with the UIMA framework using other standards as well.

O13 - Multimodal Corpora (1)

Wednesday, May 23, 18:10

Chairperson: **Christopher Cieri** Oral Session

The REPERE Corpus : a multimodal corpus for person recognition

Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert and Ludovic Quintard

The REPERE Challenge aims to support research on people recognition in multimodal conditions. To assess the technology progression, annual evaluation campaigns will be organized from 2012 to 2014. In this context, the REPERE corpus, a French videos corpus with multimodal annotation, has been developed. This paper presents datasets collected for the dry run test that took place at the beginning of 2012. Specific annotation tools and

guidelines are mainly described. At the time being, 6 hours of data have been collected and annotated. Last section presents analyses of annotation distribution and interaction between modalities in the corpus.

Polish Multimodal Corpus – a collection of referential gestures

Magdalena Lis

In face to face interaction, people refer to objects and events not only by means of speech but also by means of gesture. The present paper describes building a corpus of referential gestures. The aim is to investigate gestural reference by incorporating insights from semantic ontologies and by employing a more holistic view on referential gestures. The paper's focus is on presenting the data collection procedure and discussing the corpus' design; additionally the first insights from constructing the annotation scheme are described.

An audiovisual political speech analysis incorporating eye-tracking and perception data

Stefan Scherer, Georg Layher, John Kane, Heiko Neumann and Nick Campbell

We investigate the influence of audiovisual features on the perception of speaking style and performance of politicians, utilizing a large publicly available dataset of German parliament recordings. We conduct a human perception experiment involving eye-tracker data to evaluate human ratings as well as behavior in two separate conditions, i.e. audiovisual and video only. The ratings are evaluated on a five dimensional scale comprising measures of insecurity, monotony, expressiveness, persuasiveness, and overall performance. Further, they are statistically analyzed and put into context in a multimodal feature analysis, involving measures of prosody, voice quality and motion energy. The analysis reveals several statistically significant features, such as pause timing, voice quality measures and motion energy, that highly positively or negatively correlate with certain human ratings of speaking style. Additionally, we compare the gaze behavior of the human subjects to evaluate saliency regions in the multimodal and visual only conditions. The eye-tracking analysis reveals significant changes in the gaze behavior of the human subjects; participants reduce their focus of attention in the audiovisual condition mainly to the region of the face of the politician and scan the upper body, including hands and arms, in the video only condition.

O14 - Machine Translation and Evaluation (1)

Wednesday, May 23, 18:10

Chairperson: **Robert Frederking** Oral Session

Eye Tracking as a Tool for Machine Translation Error Analysis

Sara Stymne, Henrik Danielsson, Sofia Bremin, Hongzhan Hu, Johanna Karlsson, Anna Prytz Lillkull and Martin Wester

We present a preliminary study where we use eye tracking as a complement to machine translation (MT) error analysis, the task of identifying and classifying MT errors. We performed a user study where subjects read short texts translated by three MT systems and one human translation, while we gathered eye tracking data. The subjects were also asked comprehension questions about the text, and were asked to estimate the text quality. We found that there are a longer gaze time and a higher number of fixations on MT errors, than on correct parts. There are also differences in the gaze time of different error types, with word order errors having the longest gaze time. We also found correlations between eye tracking data and human estimates of text quality. Overall our study shows that eye tracking can give complementary information to error analysis, such as aiding in ranking error types for seriousness.

Involving Language Professionals in the Evaluation of Machine Translation

Eleftherios Avramidis, Aljoscha Burchardt, Christian Federmann, Maja Popović, Cindy Tscherwinka and David Vilar

Significant breakthroughs in machine translation only seem possible if human translators are taken into the loop. While automatic evaluation and scoring mechanisms such as BLEU have enabled the fast development of systems, it is not clear how systems can meet real-world (quality) requirements in industrial translation scenarios today. The taraXÜ project paves the way for wide usage of hybrid machine translation outputs through various feedback loops in system development. In a consortium of research and industry partners, the project integrates human translators into the development process for rating and post-editing of machine translation outputs thus collecting feedback for possible improvements.

An Analysis (and an Annotated Corpus) of User Responses to Machine Translation Output

Daniele Pighin, Lluís Màrquez and Jonathan May

We present an annotated resource consisting of open-domain translation requests, automatic translations and user-provided

corrections collected from casual users of the translation portal http://reverso.net. The layers of annotation provide: 1) quality assessments for 830 correction suggestions for translations into English, at the segment level, and 2) 814 usefulness assessments for English-Spanish and English-French translation suggestions, a suggestion being useful if it contains at least local clues that can be used to improve translation quality. We also discuss the results of our preliminary experiments concerning 1) the development of an automatic filter to separate useful from non-useful feedback, and 2) the incorporation in the machine translation pipeline of bilingual phrases extracted from the suggestions. The annotated data, available for download from ftp://mi.eng.cam.ac.uk/data/faust/LW-UPC-Oct11-FAUST-feedback-annotation.tgz, is released under a Creative Commons license. To our best knowledge, this is the first resource of this kind that has ever been made publicly available.

O15 - Information Extraction and Question Answering

Wednesday, May 23, 18:10

Chairperson: **Allan Hanbury** Oral Session

Challenges in the Knowledge Base Population Slot Filling Task

Bonan Min and Ralph Grishman

The Knowledge Based Population (KBP) evaluation track of the Text Analysis Conferences (TAC) has been held for the past 3 years. One of the two tasks of KBP is slot filling: finding within a large corpus the values of a set of attributes of given people and organizations. This task has proven very challenging, with top systems rarely exceeding 30% F-measure. In this paper, we present an error analysis and classification for those answers which could be found by a manual corpus search but were not found by any of the systems participating in the 2010 evaluation. The most common sources of failure were limitations on inference, errors in coreference (particularly with nominal anaphors), and errors in named entity recognition. We relate the types of errors to the characteristics of the task and show the wide diversity of problems that must be addressed to improve overall performance.

Evaluating Machine Reading Systems through Comprehension Tests

Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Richard Sutcliffe, Corina Forascu and Caroline Sporleder

This paper describes a methodology for testing and evaluating the performance of Machine Reading systems through Question Answering and Reading Comprehension Tests. The methodology is being used in QA4MRE (QA for Machine Reading Evaluation), one of the labs of CLEF. The task was to answer a series of multiple choice tests, each based on a single document. This allows complex questions to be asked but makes evaluation simple and completely automatic. The evaluation architecture is completely multilingual: test documents, questions, and their answers are identical in all the supported languages. Background text collections are comparable collections harvested from the web for a set of predefined topics. Each test received an evaluation score between 0 and 1 using c@1. This measure encourages systems to reduce the number of incorrect answers while maintaining the number of correct ones by leaving some questions unanswered. 12 groups participated in the task, submitting 62 runs in 3 different languages (German, English, and Romanian). All runs were monolingual; no team attempted a cross-language task. We report here the conclusions and lessons learned after the first campaign in 2011.

Biomedical Chinese-English CLIR Using an Extended CMeSH Resource to Expand Queries

Xinkai Wang, Paul Thompson, Jun'ichi Tsujii and Sophia Ananiadou

Cross-lingual information retrieval (CLIR) involving the Chinese language has been thoroughly studied in the general language domain, but rarely in the biomedical domain, due to the lack of suitable linguistic resources and parsing tools. In this paper, we describe a Chinese-English CLIR system for biomedical literature, which exploits a bilingual ontology, the "eCMeSH Tree". This is an extension of the Chinese Medical Subject Headings (CMeSH) Tree, based on Medical Subject Headings (MeSH). Using the 2006 and 2007 TREC Genomics track data, we have evaluated the performance of the eCMeSH Tree in expanding queries. We have compared our results to those obtained using two other approaches, i.e. pseudo-relevance feedback (PRF) and document translation (DT). Subsequently, we evaluate the performance of different combinations of these three retrieval methods. Our results show that our method of expanding queries using the eCMeSH Tree can outperform the PRF method. Furthermore, combining this method with PRF and DT helps to smooth the differences in query expansion, and consequently results in the best performance amongst all experiments reported. All experiments compare the use of two different retrieval models, i.e. Okapi BM25 and a query likelihood language model. In general, the former performs slightly better.

O16 - Web Services

Wednesday, May 23, 18:10

Chairperson: **Marc Kemps-Snijders** Oral Session

Towards a User-Friendly Platform for Building Language Resources based on Web Services

Marc Poch, Antonio Toral, Olivier Hamon, Valeria Quochi and Núria Bel

This paper presents the platform developed in the PANACEA project, a distributed factory that automates the stages involved in the acquisition, production, updating and maintenance of Language Resources required by Machine Translation and other Language Technologies. We adopt a set of tools that have been successfully used in the Bioinformatics field, they are adapted to the needs of our field and used to deploy web services, which can be combined to build more complex processing chains (workflows). This paper describes the platform and its different components (web services, registry, workflows, social network and interoperability). We demonstrate the scalability of the platform by carrying out a set of massive data experiments. Finally, a validation of the platform across a set of required criteria proves its usability for different types of users (non-technical users and providers).

Web Service integration platform for Polish linguistic resources

Maciej Ogrodniczuk and Michał Lenart

This paper presents a robust linguistic Web service framework for Polish, combining several mature offline linguistic tools in a common online platform. The toolset comprise paragraph-, sentence- and token-level segmenter, morphological analyser, disambiguating tagger, shallow and deep parser, named entity recognizer and coreference resolver. Uniform access to processing results is provided by means of a stand-off packaged adaptation of National Corpus of Polish TEI P5-based representation and interchange format. A concept of asynchronous handling of requests sent to the implemented Web service (Multiservice) is introduced to enable processing large amounts of text by setting up language processing chains of desired complexity. Apart from a dedicated API, a simpleWeb interface to the service is presented, allowing to compose a chain of annotation services, run it and periodically check for execution results, made available as plain XML or in a simple visualization. Usage examples and results from performance and scalability tests are also included.

Classifying Standard Linguistic Processing Functionalities based on Fundamental Data Operation Types

Yoshihiko Hayashi and Chiharu Narawa

iIt is often argued that a set of standard linguistic processing functionalities should be identified,with each of them given a formal specification. We would benefit from the formal specifications; for example, the semi-automated composition of a complex language processing workflow could be enabled in due time. This paper extracts a standard set of linguistic processing functionalities and tries to classify them formally. To do this, we first investigated prominent types of language Web services/linguistic processors by surveying a Web-based language service infrastructure and published NLP toolkits. We next induced a set of standard linguistic processing functionalities by carefully investigating each of the linguistic processor types. The standard linguistic processing functionalities was then characterized by the input/output data types, as well as the required data operation types, which were also derived from the investigation. As a result, we came up with an ontological depiction that classifies linguistic processors and linguistic processing functionalities with respect to the fundamental data operation types. We argue that such an ontological depiction can explicitly describe the functional aspects of a linguistic processing functionality.

P12 - Subjectivity: Sentiments, Emotions, Opinions (1)

Wednesday, May 23, 18:10-19:30

Chairperson: **Carlo Strapparava** Poster Session

A Multilingual Natural Stress Emotion Database

Xin Zuo, Tian Li and Pascale Fung

In this paper, we describe an ongoing effort in collecting and annotating a multilingual speech database of natural stress emotion from university students. The goal is to detect natural stress emotions and study the stress expression differences in different languages, which may help psychologists in the future. We designed a common questionnaire of stress-inducing and non-stress-inducing questions in English, Mandarin and Cantonese and collected a first ever, multilingual corpus of natural stress emotion. All of the students are native speakers of the corresponding language. We asked native language speakers to annotate recordings according to the participants' self-label states and obtained a very good kappa inter labeler agreement. We carried out human perception tests where listeners who do not understand Chinese were asked to detect stress emotion from the Mandarin

Chinese database. Compared to the annotation labels, these human perceived emotions are of low accuracy, which shows a great necessity for natural stress detection research.

Method for Collection of Acted Speech Using Various Situation Scripts

Takahiro Miyajima, Hideaki Kikuchi, Katsuhiko Shirai and Shigeki Okawa

This study was carried out to improve the quality of acted emotional speech. In the recent paradigm shift in speech collection techniques, methods for the collection of high-quality and spontaneous speech has been strongly focused on. However, such methods involve various constraints: such as the difficulty in controlling utterances and sound quality. Hence, our study daringly focuses on acted speech because of its high operability. In this paper, we propose a new method for speech collection by refining acting scripts. We compared the speech collected using our proposed method and that collected using an imitation of the legacy method that was implemented with traditional basic emotional words. The results show the advantage of our proposed method, i.e., the possibility of the generating high F0 fluctuations in acoustical expressions, which is one of the important features of the expressive speech, while ensuring that there is no decline in the naturalness and other psychological features.

Annotating Opinions in German Political News

Hong Li, Xiwen Cheng, Kristina Adson, Tal Kirshboim and Feiyu Xu

This paper presents an approach to construction of an annotated corpus for German political news for the opinion mining task. The annotated corpus has been applied to learn relation extraction rules for extraction of opinion holders, opinion content and classification of polarities. An adapted annotated schema has been developed on top of the state-of-the-art research. Furthermore, a general tool for annotating relations has been utilized for the annotation task. An evaluation of the inter-annotator agreement has been conducted. The rule learning is realized with the help of a minimally supervised machine learning framework DARE.

Hindi Subjective Lexicon: A Lexical Resource for Hindi Adjective Polarity Classification

Akshat Bakliwal, Piyush Arora and Vasudeva Varma

With recent developments in web technologies, percentage web content in Hindi is growing up at a lighting speed. This information can prove to be very useful for researchers, governments and organization to learn what's on public mind, to make sound decisions. In this paper, we present a graph based wordnet expansion method to generate a full (adjective and adverb) subjective lexicon. We used synonym and antonym relations to expand the initial seed lexicon. We show three different evaluation strategies to validate the lexicon. We achieve 70.4% agreement with human annotators and 79% accuracy on product review classification. Main contribution of our work 1) Developing a lexicon of adjectives and adverbs with polarity scores using Hindi Wordnet. 2) Developing an annotated corpora of Hindi Product Reviews.

The I3MEDIA speech database: a trilingual annotated corpus for the analysis and synthesis of emotional speech

Juan María Garrido, Yesika Laplaza, Montse Marquina, Andrea Pearman, José Gregorio Escalada, Miguel Ángel Rodríguez and Ana Armenta

In this article the I3Media corpus is presented, a trilingual (Catalan, English, Spanish) speech database of neutral and emotional material collected for analysis and synthesis purposes. The corpus is actually made up of six different subsets of material: a neutral subcorpus, containing emotionless utterances; a 'dialog' subcorpus, containing typical call center utterances; an 'emotional' corpus, a set of sentences representative of pure emotional states; a 'football' subcorpus, including utterances imitating a football broadcasting situation; a 'SMS' subcorpus, including readings of SMS texts; and a 'paralinguistic elements' corpus, including recordings of interjections and paralinguistic sounds uttered in isolation. The corpus was read by professional speakers (male, in the case of Spanish and Catalan; female, in the case of the English corpus), carefully selected to meet criteria of language competence, voice quality and acting conditions. It is the result of a collaboration between the Speech Technology Group at Telefónica Investigación y Desarrollo (TID) and the Speech and Language Group at Barcelona Media Centre d'Innovació (BM), as part of the I3Media project.

A hierarchical approach with feature selection for emotion recognition from speech

Panagiotis Giannoulis and Gerasimos Potamianos

We examine speaker independent emotion classification from speech, reporting experiments on the Berlin database across six basic emotions. Our approach is novel in a number of ways: First, it is hierarchical, motivated by our belief that the most suitable feature set for classification is different for each pair of emotions. Further, it uses a large number of feature sets of different types, such as prosodic, spectral, glottal flow based, and AM-FM ones. Finally, it employs a two-stage feature selection strategy to achieve discriminative dimensionality reduction. The approach results to a classification rate of 85%, comparable to the state-of-the-art on this dataset.

Extending the EmotiNet Knowledge Base to Improve the Automatic Detection of Implicitly Expressed Emotions from Text

Alexandra Balahur and Jesús M. Hermida

Sentiment analysis is one of the recent, highly dynamic fields in Natural Language Processing. Although much research has been performed in this area, most existing approaches are based on word-level analysis of texts and are mostly able to detect only explicit expressions of sentiment. However, in many cases, emotions are not expressed by using words with an affective meaning (e.g. happy), but by describing real-life situations, which readers (based on their commonsense knowledge) detect as being related to a specific emotion. Given the challenges of detecting emotions from contexts in which no lexical clue is present, in this article we present a comparative analysis between the performance of well-established methods for emotion detection (supervised and lexical knowledge-based) and a method we extend, which is based on commonsense knowledge stored in the EmotiNet knowledge base. Our extensive comparative evaluations show that, in the context of this task, the approach based on EmotiNet is the most appropriate.

Fine-grained German Sentiment Analysis on Social Media

Saeedeh Momtazi

Expressing opinions and emotions on social media becomes a frequent activity in daily life. People express their opinions about various targets via social media and they are also interested to know about other opinions on the same target. Automatically identifying the sentiment of these texts and also the strength of the opinions is an enormous help for people and organizations who are willing to use this information for their goals. In this paper, we present a rule-based approach for German sentiment analysis. The proposed model provides a fine-grained annotation for German texts, which represents the sentiment strength of the input text using two scores: positive and negative. The scores show that if the text contains any positive or negative opinion as well as the strength of each positive and negative opinions. To this aim, a German opinion dictionary of 1,864 words is prepared and compared with other opinion dictionaries for German. We also introduce a new dataset for German sentiment analysis. The dataset contains 500 short texts from social media about German celebrities and is annotated by three annotators. The results show that the proposed unsupervised model outperforms the supervised machine learning techniques. Moreover, the new dictionary performs better than other German opinion dictionaries.

"You Seem Aggressive!" Monitoring Anger in a Practical Application

Felix Burkhardt

A monitoring system to detect emotional outbursts in day-to-day communication is presented. The anger monitor was tested in a household and in parallel in an office surrounding. Although the state of the art of emotion recognition seems sufficient for practical applications, the acquisition of good training material remains a difficult task, as cross database performance is too low to be used in this context. A solution will probably consist of the combination of carefully drafted general training databases and the development of usability concepts to (re-) train the monitor in the field.

Mining Sentiment Words from Microblogs for Predicting Writer-Reader Emotion Transition

Yi-jie Tang and Hsin-Hsi Chen

The conversations between posters and repliers in microblogs form a valuable writer-reader emotion corpus. This paper adopts a log relative frequency ratio to investigate the linguistic features which affect emotion transitions, and applies the results to predict writers' and readers' emotions. A 4-class emotion transition predictor, a 2-class writer emotion predictor, and a 2-class reader emotion predictor are proposed and compared.

Bootstrapping Sentiment Labels For Unannotated Documents With Polarity PageRank

Christian Scheible and Hinrich Schütze

We present a novel graph-theoretic method for the initial annotation of high-confidence training data for bootstrapping sentiment classifiers. We estimate polarity using topic-specific PageRank. Sentiment information is propagated from an initial seed lexicon through a joint graph representation of words and documents. We report improved classification accuracies across multiple domains for the base models and the maximum entropy model bootstrapped from the PageRank annotation.

P13 - Named Entity Recognition

Wednesday, May 23, 18:10-19:30

Chairperson: **Antonio Branco** Poster Session

Learning Categories and their Instances by Contextual Features

Antje Schlaf and Robert Remus

We present a 3-step framework that learns categories and their instances from natural language text based on given training

examples. Step 1 extracts contexts of training examples as rules describing this category from text, considering part of speech, capitalization and category membership as features. Step 2 selects high quality rules using two consequent filters. The first filter is based on the number of rule occurrences, the second filter takes two non-independent characteristics into account: a rule's precision and the amount of instances it acquires. Our framework adapts the filter's threshold values to the respective category and the textual genre by automatically evaluating rule sets resulting from different filter settings and selecting the best performing rule set accordingly. Step 3 then identifies new instances of a category using the filtered rules applied within a previously proposed algorithm. We inspect the rule filters' impact on rule set quality and evaluate our framework by learning first names, last names, professions and cities from a hitherto unexplored textual genre – search engine result snippets – and achieve high precision on average.

Rembrandt - a named-entity recognition framework

Nuno Cardoso

Rembrandt is a named entity recognition system specially crafted to annotate documents by classifying named entities and ground them into unique identifiers. Rembrandt played an important role within our research over geographic IR, thus evolving into a more capable framework where documents can be annotated, manually curated and indexed. The goal of this paper is to present Rembrandt's simple but powerful annotation framework to the NLP community.

An Adaptive Framework for Named Entity Combination

Bogdan Sacaleanu and Günter Neumann

We have developed a new OSGi-based platform for Named Entity Recognition (NER) which uses a voting strategy to combine the results produced by several existing NER systems (currently OpenNLP, LingPipe and Stanford). The different NER systems have been systematically decomposed and modularized into the same pipeline of preprocessing components in order to support a flexible selection and ordering of the NER processing flow. This high modular and component-based design supports the possibility to setup different constellations of chained processing steps including alternative voting strategies for combining the results of parallel running components.

Rule-based Entity Recognition and Coverage of SNOMED CT in Swedish Clinical Text

Maria Skeppstedt, Maria Kvist and Hercules Dalianis

Named entity recognition of the clinical entities disorders, findings and body structures is needed for information extraction from unstructured text in health records. Clinical notes from a Swedish emergency unit were annotated and used for evaluating a rule- and terminology-based entity recognition system. This system used different preprocessing techniques for matching terms to SNOMED CT, and, one by one, four other terminologies were added. For the class body structure, the results improved with preprocessing, whereas only small improvements were shown for the classes disorder and finding. The best average results were achieved when all terminologies were used together. The entity body structure was recognised with a precision of 0.74 and a recall of 0.80, whereas lower results were achieved for disorder (precision: 0.75, recall: 0.55) and for finding (precision: 0.57, recall: 0.30). The proportion of entities containing abbreviations were higher for false negatives than for correctly recognised entities, and no entities containing more than two tokens were recognised by the system. Low recall for disorders and findings shows both that additional methods are needed for entity recognition and that there are many expressions in clinical text that are not included in SNOMED CT.

Latvian and Lithuanian Named Entity Recognition with TildeNER

M$\bar{a}$rcis Pinnis

In this paper the author presents TildeNER – an open source freely available named entity recognition toolkit and the first multi-class named entity recognition system for Latvian and Lithuanian languages. The system is built upon a supervised conditional random field classifier and features heuristic and statistical refinement methods that improve supervised classification, thus boosting the overall system's performance. The toolkit provides means for named entity recognition model bootstrapping, plaintext document and also pre-processed (morpho-syntactically tagged) tab-separated document named entity tagging and evaluation on test data. The paper presents the design of the system, describes the most important data formats and briefly discusses extension possibilities to different languages. It also gives evaluation on human annotated gold standard test corpora for Latvian and Lithuanian languages as well as comparative performance analysis to a state-of-the art English named entity recognition system using parallel and strongly comparable corpora. The author gives analysis of the Latvian and Lithuanian named entity tagged corpora annotation process and the created named entity annotated corpora.

Tree-Structured Named Entity Recognition on OCR Data: Analysis, Processing and Results

Marco Dinarelli and Sophie Rosset

In this paper we deal with named entity detection on data acquired via OCR process on documents dating from 1890. The resulting

corpus is very noisy. We perform an analysis to find possible strategies to overcome errors introduced by the OCR process. We propose a preprocessing procedure in three steps to clean data and correct, at least in part, OCR mistakes. The task is made even harder by the complex tree-structure of named entities annotated on data, we solve this problem however by adopting an effective named entity detection system we proposed in previous work. We evaluate our procedure for preprocessing OCR-ized data in two ways: in terms of perplexity and OOV rate of a language model on development and evaluation data, and in terms of the performance of the named entity detection system on the preprocessed data. The preprocessing procedure results to be effective, allowing to improve by a large margin the system we proposed for the official evaluation campaign on Old Press, and allowing to outperform also the best performing system of the evaluation campaign.

Aleda, a free large-scale entity database for French

Benoît Sagot and Rosa Stern

Named entity recognition, which focuses on the identification of the span and type of named entity mentions in texts, has drawn the attention of the NLP community for a long time. However, many real-life applications need to know which real entity each mention refers to. For such a purpose, often refered to as entity resolution and linking, an inventory of entities is required in order to constitute a reference. In this paper, we describe how we extracted such a resource for French from freely available resources (the French Wikipedia and the GeoNames database). We describe the results of an instrinsic evaluation of the resulting entity database, named Aleda, as well as those of a task-based evaluation in the context of a named entity detection system. We also compare it with the NLGbAse database (Charton and Torres-Moreno, 2010), a resource with similar objectives.

Evaluating the Impact of Phrase Recognition on Concept Tagging

Pablo Mendes, Joachim Daiber, Rohana Rajapakse, Felix Sasaki and Christian Bizer

We have developed DBpedia Spotlight, a flexible concept tagging system that is able to annotate entities, topics and other terms in natural language text. The system starts by recognizing phrases to annotate in the input text, and subsequently disambiguates them to a reference knowledge base extracted from Wikipedia. In this paper we evaluate the impact of the phrase recognition step on the ability of the system to correctly reproduce the annotations of a gold standard in an unsupervised setting. We argue that a combination of techniques is needed, and we evaluate a number of alternatives according to an existing evaluation set.

P14 - Dialogue

Wednesday, May 23, 18:10-19:30

Chairperson: **Ron Artstein** Poster Session

Adaptive Speech Understanding for Intuitive Model-based Spoken Dialogues

Tobias Heinroth, Maximilian Grotz, Florian Nothdurft and Wolfgang Minker

In this paper we present three approaches towards adaptive speech understanding. The target system is a model-based Adaptive Spoken Dialogue Manager, the OwlSpeak ASDM. We enhanced this system in order to properly react on non-understandings in real-life situations where intuitive communication is required. OwlSpeak provides a model-based spoken interface to an Intelligent Environment depending on and adapting to the current context. It utilises a set of ontologies used as dialogue models that can be combined dynamically during runtime. Besides the benefits the system showed in practice, real-life evaluations also conveyed some limitations of the model-based approach. Since it is unfeasible to model all variations of the communication between the user and the system beforehand, various situations where the system did not correctly understand the user input have been observed. Thus we present three enhancements towards a more sophisticated use of the ontology-based dialogue models and show how grammars may dynamically be adapted in order to understand intuitive user utterances. The evaluation of our approaches revealed the incorporation of a lexical-semantic knowledgebase into the recognition process to be the most promising approach.

Relating Dominance of Dialogue Participants with their Verbal Intelligence Scores

Kseniya Zablotskaya, Umair Rahim, Fernando Fernández Martínez and Wolfgang Minker

In this work we investigated whether there is a relationship between dominant behaviour of dialogue participants and their verbal intelligence. The analysis is based on a corpus containing 56 dialogues and verbal intelligence scores of the test persons. All the dialogues were divided into three groups: H-H is a group of dialogues between higher verbal intelligence participants, L-L is a group of dialogues between lower verbal intelligence participant and L-H is a group of all the other dialogues. The dominance scores of the dialogue partners from each group were analysed. The analysis showed that differences between

dominance scores and verbal intelligence coefficients for L-L were positively correlated. Verbal intelligence scores of the test persons were compared to other features that may reflect dominant behaviour. The analysis showed that number of interruptions, long utterances, times grabbed the floor, influence diffusion model, number of agreements and several acoustic features may be related to verbal intelligence. These features were used for the automatic classification of the dialogue partners into two groups (lower and higher verbal intelligence participants); the achieved accuracy was 89.36%.

The coding and annotation of multimodal dialogue acts

Volha Petukhova and Harry Bunt

Recent years have witnessed a growing interest in annotating linguistic data at the semantic level, including the annotation of dialogue corpus data. The annotation scheme developed as international standard for dialogue act annotation ISO 24617-2 is based on the DIT++ scheme (Bunt, 2006; 2009) which combines the multidimensional DIT scheme (Bunt, 1994) with concepts from DAMSL (Allen and Core , 1997) and various other schemes. This scheme is designed in a such way that it can be applied not only to spoken dialogue, as is the case for most of the previously defined dialogue annotation schemes, but also to multimodal dialogue. This paper describes how the ISO 24617-2 annotation scheme can be used, together with the DIT++ method of 'multidimensional segmentation', to annotate nonverbal and multimodal dialogue behaviour. We analyse the fundamental distinction between (a) the coding of surface features; (b) form-related semantic classification; and (c) semantic annotation in terms of dialogue acts, supported by experimental studies of (a) and (b). We discuss examples of specification languages for representing the results of each of these activities, show how dialogue act annotations can be attached to XML representations of functional segments of multimodal data.

Using DiAML and ANVIL for multimodal dialogue annotations

Harry Bunt, Michael Kipp and Volha Petukhova

This paper shows how interoperable dialogue act annotations, using the multidimensional annotation scheme and the markup language DiAML of ISO standard 24617-2, can conveniently be obtained using the newly implemented facility in the ANVIL annotation tool to produce XML-based output directly in the DiAML format. ANVIL offers the use of multiple user-defined 'tiers' for annotating various kinds of information. This is shown to be convenient not only for multimodal information but also for dialogue act annotation according to ISO standard 24617-2 because of the latter's multidimensionality: functional dialogue segments are viewed as expressing one or more dialogue acts, and every dialogue act belongs to one of a number of dimensions of communication, defined in the standard, for each of which a different ANVIL tier can conveniently be used. Annotations made in the multi-tier interface can be exported in the ISO 24617-2 format, thus supporting the creation of interoperable annotated corpora of multimodal dialogue.

A Scalable Architecture For Web Deployment of Spoken Dialogue Systems

Matthew Fuchs, Nikos Tsourakis and Manny Rayner

We describe a scalable architecture, particularly well-suited to cloud-based computing, which can be used for Web-deployment of spoken dialogue systems. In common with similar platforms, like WAMI and the Nuance Mobile Developer Platform, we use a client/server approach in which speech recognition is carried out on the server side; our architecture, however, differs from these systems in offering considerably more elaborate server-side functionality, based on large-scale grammar-based language processing and generic dialogue management. We describe two substantial applications, built using our framework, which we argue would have been hard to construct in WAMI or NMDP. Finally, we present a series of evaluations carried out using CALL-SLT, a speech translation game, where we contrast performance in Web and desktop versions. Task Error Rate in the Web version is only slightly inferior that in the desktop one, and the average additional latency is under half a second. The software is generally available for research purposes.

A Corpus for a Gesture-Controlled Mobile Spoken Dialogue System

Nikos Tsourakis and Manny Rayner

Speech and hand gestures offer the most natural modalities for everyday human-to-human interaction. The availability of diverse spoken dialogue applications and the proliferation of accelerometers on consumer electronics allow the introduction of new interaction paradigms based on speech and gestures. Little attention has been paid however to the manipulation of spoken dialogue systems through gestures. Situation-induced disabilities or real disabilities are determinant factors that motivate this type of interaction. In this paper we propose six concise and intuitively meaningful gestures that can be used to trigger the commands in any SDS. Using different machine learning techniques we achieve a classification error for the gesture patterns of less than 5%, and we also compare our own set of gestures to ones proposed by users. Finally, we examine the social acceptability of the specific interaction scheme and encounter high levels of acceptance for public use.

A Corpus of Spontaneous Multi-party Conversation in Bosnian Serbo-Croatian and British English

Emina Kurtic, Bill Wells, Guy J. Brown, Timothy Kempton and Ahmet Aker

In this paper we present a corpus of audio and video recordings of spontaneous, face-to-face multi-party conversation in two languages. Freely available high quality recordings of mundane, non-institutional, multi-party talk are still sparse, and this corpus aims to contribute valuable data suitable for study of multiple aspects of spoken interaction. In particular, it constitutes a unique resource for spoken Bosnian Serbo-Croatian (BSC), an under-resourced language with no spoken resources available at present. The corpus consists of just over 3 hours of free conversation in each of the target languages, BSC and British English (BE). The audio recordings have been made on separate channels using head-set microphones, as well as using a microphone array, containing 8 omni-directional microphones. The data has been segmented and transcribed using segmentation notions and transcription conventions developed from those of the conversation analysis research tradition. Furthermore, the transcriptions have been automatically aligned with the audio at the word and phone level, using the method of forced alignment. In this paper we describe the procedures behind the corpus creation and present the main features of the corpus for the study of conversation.

The Herme Database of Spontaneous Multimodal Human-Robot Dialogues

Jing Guang Han, Emer Gilmartin, Celine DeLooze, Brian Vaughan and Nick Campbell

This paper presents methodologies and tools for language resource (LR) construction. It describes a database of interactive speech collected over a three-month period at the Science Gallery in Dublin, where visitors could take part in a conversation with a robot. The system collected samples of informal, chatty dialogue – normally difficult to capture under laboratory conditions for human-human dialogue, and particularly so for human-machine interaction. The conversations were based on a script followed by the robot consisting largely of social chat with some task-based elements. The interactions were audio-visually recorded using several cameras together with microphones. As part of the conversation the participants were asked to sign a consent form giving permission to use their data for human-machine interaction research. The multimodal corpus will be made available to interested researchers and the technology developed during the three-month exhibition is being extended for use in education and assisted-living applications.

Annotation of response tokens and their triggering expressions in Japanese multi-party conversations

Yasuharu Den, Hanae Koiso, Katsuya Takanashi and Nao Yoshida

In this paper, we propose a new scheme for annotating response tokens (RTs) and their triggering expressions in Japanese multi-party conversations. In the proposed scheme, RTs are first identified and classified according to their forms, and then sub-classified according to their sequential positions in the discourse. To deeply study the contexts in which RTs are used, the scheme also provides procedures for annotating triggering expressions, which are considered to trigger the listener's production of RTs. RTs are classified according to whether or not there is a particular object or proposition in the speaker's turn for which the listener shows a positive or aligned stance. Triggering expressions are then identified in the speaker's turn; they include surprising facts and other newsworthy things, opinions and assessments, focus of a response to a question or repair initiation, keywords in narratives, and embedded propositions quoted from other's statement or thought, which are to be agreed upon, assessed, or noticed. As an illustrative application of our scheme, we present a preliminary analysis on the distribution of the latency of the listener's response to the triggering expression, showing how it differs according to RT's forms and positions.

Syntactic annotation of spontaneous speech: application to call-center conversation data

Thierry Bazillon, Melanie Deplano, Frederic Bechet, Alexis Nasr and Benoit Favre

This paper describes the syntactic annotation process of the DECODA corpus. This corpus contains manual transcriptions of spoken conversations recorded in the French call-center of the Paris Public Transport Authority (RATP). Three levels of syntactic annotation have been performed with a semi-supervised approach: POS tags, Syntactic Chunks and Dependency parses. The main idea is to use off-the-shelf NLP tools and models, originaly developped and trained on written text, to perform a first automatic annotation on the manually transcribed corpus. At the same time a fully manual annotation process is performed on a subset of the original corpus, called the GOLD corpus. An iterative process is then applied, consisting in manually correcting errors found in the automatic annotations, retraining the linguistic models of the NLP tools on this corrected corpus, then checking the quality of the adapted models on the fully manual annotations of the GOLD corpus. This process iterates until a certain error rate is reached. This paper describes this process, the main issues raising when adapting NLP tools to process speech transcriptions, and presents the first evaluations performed with these new adapted tools.

DECODA: a call-centre human-human spoken conversation corpus

Frederic Bechet, Benjamin Maza, Nicolas Bigouroux, Thierry Bazillon, Marc El-Beze, Renato De Mori and Eric Arbillot

The goal of the DECODA project is to reduce the development cost of Speech Analytics systems by reducing the need for manual annotat ion. This project aims to propose robust speech data mining tools in the framework of call-center monitoring and evaluation, by means of weakly supervised methods. The applicative framework of the project is the call-center of the RATP (Paris public transport authority). This project tackles two very important open issues in the development of speech mining methods from spontaneous speech recorded in call-centers : robustness (how to extract relevant information from very noisy and spontaneous speech messages) and weak supervision (how to reduce the annotation effort needed to train and adapt recognition and classification models). This paper describes the DECODA corpus collected at the RATP during the project. We present the different annotation levels performed on the corpus, the methods used to obtain them, as well as some evaluation of the quality of the annotations produced.

Resource Evaluation for Usable Speech Interfaces: Utilizing Human–Human Dialogue

Pepi Stavropoulou, Dimitris Spiliotopoulos and Georgios Kouroupetroglou

Human-human spoken dialogues are considered an important tool for effective speech interface design and are often used for stochastic model training in speech based applications. However, the less restricted nature of human-human interaction compared to human-system interaction may undermine the usefulness of such corpora for creating effective and usable interfaces. In this respect, this work examines the differences between corpora collected from human-human interaction and corpora collected from actual system use, in order to formally assess the appropriateness of the former for both the design and implementation of spoken dialogue systems. Comparison results show that there are significant differences with respect to vocabulary, sentence structure and speech recognition success rate among others. Nevertheless, compared to other available tools and techniques, human-human dialogues may still be used as a temporary at least solution for building more effective working systems. Accordingly, ways to better utilize such resources are presented.

3rd party observer gaze as a continuous measure of dialogue flow

Jens Edlund, Simon Alexandersson, Jonas Beskow, Lisa Gustavsson, Mattias Heldner, Anna Hjalmarsson, Petter Kallionen and Ellen Marklund

We present an attempt at using 3rd party observer gaze to get a measure of how appropriate each segment in a dialogue is for a speaker change. The method is a step away from the current dependency of speaker turns or talkspurts towards a more general view of speaker changes. We show that 3rd party observers do indeed largely look at the same thing (the speaker), and how this can be captured and utilized to provide insights into human communication. In addition, the results also suggest that there might be differences in the distribution of 3rd party observer gaze depending on how information-rich an utterance is.

Pursing power in Arabic on-line discussion forums

Marc Tomlinson, David Bracewell, Mary Draper, Zewar Almissour, Ying Shi and Jeremy Bensley

We present a novel corpus for identifying individuals within a group setting that are attempting to gain power within the group. The corpus is entirely in Arabic and is derived from the on-line WikiTalk discussion forums. Entries on the forums were annotated at multiple levels, top-level annotations identified whether an individual was pursuing power on the forum, and low level annotations identified linguistic indicators that signaled an individuals social intentions. An analysis of our annotations reflects a high-degree of overlap between current theories on power and conflict within a group and the behavior of individuals within the transcripts. The described datasource provides an appropriate means for modeling an individual's pursuit of power within an on-line discussion group and also allows for enumeration and validation of current theories on the ways in which individuals strive for power.

Causal analysis of task completion errors in spoken music retrieval interactions

Sunao Hara, Norihide Kitaoka and Kazuya Takeda

In this paper, we analyze the causes of task completion errors in spoken dialog systems, using a decision tree with N-gram features of the dialog to detect task-incomplete dialogs. The dialog for a music retrieval task is described by a sequence of tags related to user and system utterances and behaviors. The dialogs are manually classified into two classes: completed and uncompleted music retrieval tasks. Differences in tag classification performance between the two classes are discussed. We then construct decision trees which can detect if a dialog

finished with the task completed or not, using information gain criterion. Decision trees using N-grams of manual tags and automatic tags achieved 74.2% and 80.4% classification accuracy, respectively, while the tree using interaction parameters achieved an accuracy rate of 65.7%. We also discuss more details of the causality of task incompletion for spoken dialog systems using such trees.

An Annotated Corpus of Film Dialogue for Learning and Characterizing Character Style

Marilyn Walker, Grace Lin and Jennifer Sawyer

Interactive story systems often involve dialogue with virtual dramatic characters. However, to date most character dialogue is written by hand. One way to ease the authoring process is to (semi-)automatically generate dialogue based on film characters. We extract features from dialogue of film characters in leading roles. Then we use these character-based features to drive our language generator to produce interesting utterances. This paper describes a corpus of film dialogue that we have collected from the IMSDb archive and annotated for linguistic structures and character archetypes. We extract different sets of features using external sources such as LIWC and SentiWordNet as well as using our own written scripts. The automation of feature extraction also eases the process of acquiring additional film scripts. We briefly show how film characters can be represented by models learned from the corpus, how the models can be distinguished based on different categories such as gender and film genre, and how they can be applied to a language generator to generate utterances that can be perceived as being similar to the intended character model.

Keynote Speech 1

Thursday, May 24, 9:00

Chairperson: **Jan Odijk**

The Web of Data: Decentralized, collaborative, interlinked and interoperable

Sören Auer and Sebastian Hellmann

Recently the publishing and integration of structured data on the Web gained traction with initiatives such as Linked Data, RDFa and schema.org. In this article we outline some fundamental principles and aspects of the emerging Web of Data. We stress the importance of open licenses as an enabler for collaboration, sharing and reuse of structured data on the Web. We discuss some features of the RDF data model and its suitability for integrating structured data on the Web. Two particularly crucial aspects are performance and scalability as well as conceptual interoperability, when using the Web as a medium for data integration. Last but not least we outline our vision of a Web of interlinked linguistic resources, which includes the establishment of a distributed ecosystem of heterogeneous NLP tools and services by means of structural, conceptual and access interoperability employing background knowledge from the Web of Data.

O17 - Infrastructures and Strategies for LRs (2)

Thursday, May 24, 9:45

Chairperson: **Andrejs Vasiljevs** Oral Session

The FLaReNet Strategic Language Resource Agenda

Claudia Soria, Núria Bel, Khalid Choukri, Joseph Mariani, Monica Monachini, Jan Odijk, Stelios Piperidis, Valeria Quochi and Nicoletta Calzolari

The FLaReNet Strategic Agenda highlights the most pressing needs for the sector of Language Resources and Technologies and presents a set of recommendations for its development and progress in Europe, as issued from a three-year consultation of the FLaReNet European project. The FLaReNet recommendations are organised around nine dimensions: a) documentation b) interoperability c) availability, sharing and distribution d) coverage, quality and adequacy e) sustainability f) recognition g) development h) infrastructure and i) international cooperation. As such, they cover a broad range of topics and activities, spanning over production and use of language resources, licensing, maintenance and preservation issues, infrastructures for language resources, resource identification and sharing, evaluation and validation, interoperability and policy issues. The intended recipients belong to a large set of players and stakeholders in Language Resources and Technology, ranging from individuals to research and education institutions, to policy-makers, funding agencies, SMEs and large companies, service and media providers. The main goal of these recommendations is to serve as an instrument to support stakeholders in planning for and addressing the urgencies of the Language Resources and Technologies of the future.

Standardizing a Component Metadata Infrastructure

Daan Broeder, Dieter van Uytvanck, Maria Gavrilidou, Thorsten Trippel and Menzo Windhouwer

This paper describes the status of the standardization efforts of a Component Metadata approach for describing Language Resources with metadata. Different linguistic and Language

& Technology communities as CLARIN, META-SHARE and NaLiDa use this component approach and see its standardization of as a matter for cooperation that has the possibility to create a large interoperable domain of joint metadata. Starting with an overview of the component metadata approach together with the related semantic interoperability tools and services as the ISOcat data category registry and the relation registry we explain the standardization plan and efforts for component metadata within ISO TC37/SC4. Finally, we present information about uptake and plans of the use of component metadata within the three mentioned linguistic and L&T communities.

Citing on-line Language Resources

Daan Broeder, Dieter van Uytvanck and Gunter Senft

Although the possibility of referring or citing on-line data from publications is seen at least theoretically as an important means to provide immediate testable proof or simple illustration of a line of reasoning, the practice has not been wide-spread yet and no extensive experience has been gained about the possibilities and problems of referring to raw data-sets. This paper makes a case to investigate the possibility and need of persistent data visualization services that facilitate the inspection and evaluation of the cited data.

An Analytical Model of Language Resource Sustainability

Khalid Choukri and Victoria Arranz

This paper elaborates on a sustainability model for Language Resources, both at a descriptive and analytical level. The first part, devoted to the descriptive model, elaborates on the definition of this concept both from a general point of view and from the Human Language Technology and Language Resources perspective. The paper also intends to list an exhaustive number of factors that have an impact on this sustainability. These factors will be clustered into Pillars so as ease understanding as well as the prediction of LR sustainability itself. Rather than simply identifying a set of LRs that have been in use for a while and that one can consider as sustainable, the paper aims at first clarifying and (re)defining the concept of sustainability by also connecting it to other domains. Then it also presents a detailed decomposition of all dimensions of Language Resource features that can contribute and/or have an impact on such sustainability. Such analysis will also help anticipate and forecast sustainability for a LR before taking any decisions concerning design and production.

On Using Linked Data for Language Resource Sharing in the Long Tail of the Localisation Market

David Lewis, Alexander O'Connor, Andrzej Zydroń, Gerd Sjögren and Rahzeb Choudhury

Innovations in localisation have focused on the collection and leverage of language resources. However, smaller localisation clients and Language Service Providers are poorly positioned to exploit the benefits of language resource reuse in comparison to larger companies. Their low throughput of localised content means they have little opportunity to amass significant resources, such as Translation memories and Terminology databases, to reuse between jobs or to train statistical machine translation engines tailored to their domain specialisms and language pairs. We propose addressing this disadvantage via the sharing and pooling of language resources. However, the current localisation standards do not support multiparty sharing, are not well integrated with emerging language resource standards and do not address key requirements in determining ownership and license terms for resources. We survey standards and research in the area of Localisation, Language Resources and Language Technologies to leverage existing localisation standards via Linked Data methodologies. This points to the potential of using semantic representation of existing data models for localisation workflow metadata, terminology, parallel text, provenance and access control, which we illustrate with an RDF example.

O18 - Dialogue

Thursday, May 24, 9:45

Chairperson: **Linne Ha** Oral Session

Evaluation of Online Dialogue Policy Learning Techniques

Alexandros Papangelis, Vangelis Karkaletsis and Fillia Makedon

The number of applied Dialogue Systems is ever increasing in several service providing and other applications as a way to efficiently and inexpensively serve large numbers of customers. A DS that employs some form of adaptation to the environment and its users is called an Adaptive Dialogue System (ADS). A significant part of the research community has lately focused on ADS and many existing or novel techniques are being applied to this problem. One of the most promising techniques is Reinforcement Learning (RL) and especially online RL. This paper focuses on online RL techniques used to achieve adaptation in Dialogue Management and provides an evaluation of various such methods in an effort to aid the designers of ADS in deciding

which method to use. To the best of our knowledge there is no other work to compare online RL techniques on the dialogue management problem.

The acquisition and dialog act labeling of the EDECAN-SPORTS corpus

Lluis-F. Hurtado, Fernando Garcia, Emilio Sanchis and Encarna Segarra

In this paper, we present the acquisition and labeling processes of the EDECAN-SPORTS corpus, which is a corpus that is oriented to the development of multimodal dialog systems acquired in Spanish and Catalan. Two Wizards of Oz were used in order to better simulate the behavior of an actual system in terms of both the information used by the different modules and the communication mechanisms between these modules. User and system dialog-act labeling, as well as other information, have been obtained automatically using this acquisition method Some preliminary experimental results with the acquired corpus show the appropriateness of the proposed acquisition method for the development of dialog systems

Developing and evaluating an emergency scenario dialogue corpus

Jolanta Bachan

The present paper describes the development and evaluation of the Polish emergency dialogue corpus recorded for studying alignment phenomena in stress scenarios. The challenge is that emergency dialogues are more complex on many levels than standard information negotiation dialogues, different resources are needed for differential investigation, and resources for this kind of corpus are rare. Currently there is no comparable corpus for Polish. In the present context, alignment is meant as adaptation on the syntactic, semantic and pragmatic levels of communication between the two interlocutors, including choice of similar lexical items and speaking style. Four different dialogue scenarios were arranged and prompt speech material was created. Two maps for the map-tasks and one emergency diapix were design to prompt semi-spontaneous dialogues simulating stress and natural communicative situations. The dialogue corpus was recorded taking into account the public character of conversations in the emergency setting. The linguistic study of alignment in this kind of dialogue made it possible to design and implement a prototype of a Polish adaptive dialogue system to support stress scenario communication (not described in this paper).

Building and Exploiting a Corpus of Dialog Interactions between French Speaking Virtual and Human Agents

Lina M. Rojas-Barahona, Alejandra Lorenzo and Claire Gardent

We describe the acquisition of a dialog corpus for French based on multi-task human-machine interactions in a serious game setting. We present a tool for data collection that is configurable for multiple games; describe the data collected using this tool and the annotation schema used to annotate it; and report on the results obtained when training a classifier on the annotated data to associate each player turn with a dialog move usable by a rule based dialog manager. The collected data consists of approximately 1250 dialogs, 10454 utterances and 168509 words and will be made freely available to academic and nonprofit research.

Leveraging study of robustness and portability of spoken language understanding systems across languages and domains: the PORTMEDIA corpora

Fabrice Lefèvre, Djamel Mostefa, Laurent Besacier, Yannick Estève, Matthieu Quignard, Nathalie Camelin, Benoit Favre, Bassam Jabaian and Lina M. Rojas-Barahona

The PORTMEDIA project is intended to develop new corpora for the evaluation of spoken language understanding systems. The newly collected data are in the field of human-machine dialogue systems for tourist information in French in line with the MEDIA corpus. Transcriptions and semantic annotations, obtained by low-cost procedures, are provided to allow a thorough evaluation of the systems' capabilities in terms of robustness and portability across languages and domains. A new test set with some adaptation data is prepared for each case: in Italian as an example of a new language, for ticket reservation as an example of a new domain. Finally the work is complemented by the proposition of a new high level semantic annotation scheme well-suited to dialogue data.

O19 - Resource Creation and Acquisition

Thursday, May 24, 9:45

Chairperson: **James Pustejovsky** Oral Session

Building a Basque-Chinese Dictionary by Using English as Pivot

Xabier Saralegi, Iker Manterola and Iñaki San Vicente

Bilingual dictionaries are key resources in several fields such as translation, language learning or various NLP tasks. However,

only major languages have such resources. Automatically built dictionaries by using pivot languages could be a useful resource in these circumstances. Pivot-based bilingual dictionary building is based on merging two bilingual dictionaries which share a common language (e.g. LA-LB, LB-LC) in order to create a dictionary for a new language pair (e.g LA-LC). This process may include wrong translations due to the polisemy of words. We built Basque-Chinese (Mandarin) dictionaries automatically from Basque-English and Chinese-English dictionaries. In order to prune wrong translations we used different methods adequate for less resourced languages. Inverse Consultation and Distributional Similarity methods are used because they just depend on easily available resources. Finally, we evaluated manually the quality of the built dictionaries and the adequacy of the methods. Both Inverse Consultation and Distributional Similarity provide good precision of translations but recall is seriously damaged. Distributional similarity prunes rare translations more accurately than other methods.

Automatic lexical semantic classification of nouns

Núria Bel, Lauren Romeo and Muntsa Padró

The work we present here addresses cue-based noun classification in English and Spanish. Its main objective is to automatically acquire lexical semantic information by classifying nouns into previously known noun lexical classes. This is achieved by using particular aspects of linguistic contexts as cues that identify a specific lexical class. Here we concentrate on the task of identifying such cues and the theoretical background that allows for an assessment of the complexity of the task. The results show that, despite of the a-priori complexity of the task, cue-based classification is a useful tool in the automatic acquisition of lexical semantic classes.

Assessing Crowdsourcing Quality through Objective Tasks

Ahmet Aker, Mahmoud El-Haj, M-Dyaa Albakour and Udo Kruschwitz

The emergence of crowdsourcing as a commonly used approach to collect vast quantities of human assessments on a variety of tasks represents nothing less than a paradigm shift. This is particularly true in academic research where it has suddenly become possible to collect (high-quality) annotations rapidly without the need of an expert. In this paper we investigate factors which can influence the quality of the results obtained through Amazon's Mechanical Turk crowdsourcing platform. We investigated the impact of different presentation methods (free text versus radio buttons), workers' base (USA versus India as the main bases of MTurk workers) and payment scale (about $4, $8 and $10 per hour) on the quality of the results. For each run we assessed the results provided by 25 workers on a set of 10 tasks. We run two different experiments using objective tasks: maths and general text questions. In both tasks the answers are unique, which eliminates the uncertainty usually present in subjective tasks, where it is not clear whether the unexpected answer is caused by a lack of worker's motivation, the worker's interpretation of the task or genuine ambiguity. In this work we present our results comparing the influence of the different factors used. One of the interesting findings is that our results do not confirm previous studies which concluded that an increase in payment attracts more noise. We also find that the country of origin only has an impact in some of the categories and only in general text questions but there is no significant difference at the top pay.

Rapid creation of large-scale corpora and frequency dictionaries

Attila Zséder, Gábor Recski, Dániel Varga and András Kornai

We describe, and make public, large-scale language resources and the toolchain used in their creation, for fifteen medium density European languages: Catalan, Czech, Croatian, Danish, Dutch, Finnish, Lithuanian, Norwegian, Polish, Portuguese, Romanian, Serbian, Slovak, Spanish, and Swedish. To make the process uniform across languages, we selected tools that are either language-independent or easily customizable for each language, and reimplemented all stages that were taking too long. To achieve processing times that are insignificant compared to the time data collection (crawling) takes, we reimplemented the standard sentence- and word-level tokenizers and created new boilerplate and near-duplicate detection algorithms. Preliminary experiments with non-European languages indicate that our methods are now applicable not just to our sample, but the entire population of digitally viable languages, with the main limiting factor being the availability of high quality stemmers.

Boosting the Coverage of a Semantic Lexicon by Automatically Extracted Event Nominalizations

Kata Gábor, Marianna Apidianaki, Benoît Sagot and Éric Villemonte de la Clergerie

In this article, we present a distributional analysis method for extracting nominalization relations from monolingual corpora. The acquisition method makes use of distributional and morphological information to select nominalization candidates. We explain how the learning is performed on a dependency annotated corpus and describe the nominalization results. Furthermore, we show how these results served to enrich an existing lexical resource, the WOLF (Wordnet Libre du Franc¸ais).

We present the techniques that we developed in order to integrate the new information into WOLF, based on both its structure and content. Finally, we evaluate the validity of the automatically obtained information and the correctness of its integration into the semantic resource. The method proved to be useful for boosting the coverage of WOLF and presents the advantage of filling verbal synsets, which are particularly difficult to handle due to the high level of verbal polysemy.

O20 - Corpus and Annotation

Thursday, May 24, 9:45

Chairperson: **Nancy Ide** Oral Session

Analyzing the Impact of Prevalence on the Evaluation of a Manual Annotation Campaign

Karën Fort, Claire François, Olivier Galibert and Maha Ghribi

This article details work aiming at evaluating the quality of the manual annotation of gene renaming couples in scientific abstracts, which generates sparse annotations. To evaluate these annotations, we compare the results obtained using the commonly advocated inter-annotator agreement coefficients such as S, κ and π , the less known R, the weighted coefficients κ ω and α as well as the F-measure and the SER. We analyze to which extent they are relevant for our data. We then study the bias introduced by prevalence by changing the way the contingency table is built. We finally propose an original way to synthesize the results by computing distances between categories, based on the produced annotations.

Corpus Annotation as a Scientific Task

Donia Scott, Rossano Barone and Rob Koeling

Annotation studies in CL are generally unscientific: they are mostly not reproducible, make use of too few (and often non-independent) annotators and use guidelines that are often something of a moving target. Additionally, the notion of 'expert annotators' invariably means only that the annotators have linguistic training. While this can be acceptable in some special contexts, it is often far from ideal. This is particularly the case when subtle judgements are required or when, as increasingly, one is making use of corpora originating from technical texts that have been produced by, and intended to be consumed by, an audience of technical experts in the field. We outline a more rigorous approach to collecting human annotations, using as our example a study designed to capture judgements on the meaning of hedge words in medical records.

Document Attrition in Web Corpora: an Exploration

Stephen Wattam, Paul Rayson and Damon Berridge

Increases in the use of web data for corpus-building, coupled with the use of specialist, single-use corpora, make for an increasing reliance on language that changes quickly, affecting the long-term validity of studies based on these methods. This 'drift' through time affects both users of open-source corpora and those attempting to interpret the results of studies based on web data. The attrition of documents online, also called link rot or document half-life, has been studied many times for the purposes of optimising search engine web crawlers, producing robust and reliable archival systems, and ensuring the integrity of distributed information stores, however, the affect that attrition has upon corpora of varying construction remains largely unknown. This paper presents a preliminary investigation into the differences in attrition rate between corpora selected using different corpus construction methods. It represents the first step in a larger longitudinal analysis, and as such presents URI-based content clues, chosen to relate to studies from other areas. The ultimate goal of this larger study is to produce a detailed enumeration of the primary biases online, and identify sampling strategies which control and minimise unwanted effects of document attrition.

A Concise Query Language with Search and Transform Operations for Corpora with Multiple Levels of Annotation

Anil Kumar Singh

The usefulness of annotated corpora is greatly increased if there is an associated tool that can allow various kinds of operations to be performed in a simple way. Different kinds of annotation frameworks and many query languages for them have been proposed, including some to deal with multiple layers of annotation. We present here an easy to learn query language for a particular kind of annotation framework based on 'threaded trees', which are somewhere between the complete order of a tree and the anarchy of a graph. Through 'typed' threads, they can allow multiple levels of annotation in the same document. Our language has a simple, intuitive and concise syntax and high expressive power. It allows not only to search for complicated patterns with short queries but also allows data manipulation and specification of arbitrary return values. Many of the commonly used tasks that otherwise require writing programs, can be performed with one or more queries. We compare the language with some others and try to evaluate it.

A New Method for Evaluating Automatically Learned Terminological Taxonomies

Paola Velardi, Roberto Navigli, Stefano Faralli and Juana Maria Ruiz-Martinez

Abstract Evaluating a taxonomy learned automatically against an existing gold standard is a very complex problem, because differences stem from the number, label, depth and ordering of the taxonomy nodes. In this paper we propose casting the problem as one of comparing two hierarchical clusters. To this end we defined a variation of the Fowlkes and Mallows measure (Fowlkes and Mallows, 1983). Our method assigns a similarity value Bhat i_ (l,r) to the learned (l) and reference (r) taxonomy for each cut i of the corresponding anonymised hierarchies, starting from the topmost nodes down to the leaf concepts. For each cut i, the two hierarchies can be seen as two clusterings Chat i_ l , Chat i_ r of the leaf concepts. We assign a prize to early similarity values, i.e. when concepts are clustered in a similar way down to the lowest taxonomy levels (close to the leaf nodes). We apply our method to the evaluation of the taxonomy learning methods put forward by Navigli et al. (2011) and Kozareva and Hovy (2010).

P15 - Semantic Annotation

Thursday, May 24, 9:45

Chairperson: **Aline Villavicencio** Poster Session

Event Nominals: Annotation Guidelines and a Manually Annotated Corpus in French

Béatrice Arnulphy, Xavier Tannier and Anne Vilnat

Within the general purpose of information extraction, detection of event descriptions is an important clue. A word refering to an event is more powerful than a single word, because it implies a location, a time, protagonists (persons, organizations…). However, if verbal designations of events are well studied and easier to detect than nominal ones, nominal designations do not claim as much definition effort and resources. In this work, we focus on nominals desribing events. As our application domain is information extraction, we follow a named entity approach to describe and annotate events. In this paper, we present a typology and annotation guidelines for event nominals annotation. We applied them to French newswire articles and produced an annotated corpus. We present observations about the designations used in our manually annotated corpus and the behavior of their triggers. We provide statistics concerning word ambiguity and context of use of event nominals, as well as machine learning experiments showing the difficulty of using lexicons for extracting events.

Building a Corpus of Indefinite Uses Annotated with Fine-grained Semantic Functions

Maria Aloni, Andreas van Cranenburgh, Raquel Fernandez and Marta Sznajder

Natural languages possess a wealth of indefinite forms that typically differ in distribution and interpretation. Although formal semanticists have strived to develop precise meaning representations for different indefinite functions, to date there has hardly been any corpus work on the topic. In this paper, we present the results of a small corpus study where English indefinite forms ‘any’ and ‘some’ were labelled with fine-grained semantic functions well-motivated by typological studies. We developed annotation guidelines that could be used by non-expert annotators and calculated inter-annotator agreement amongst several coders. The results show that the annotation task is hard, with agreement scores ranging from 52% to 62% depending on the number of functions considered, but also that each of the independent annotations is in accordance with theoretical predictions regarding the possible distributions of indefinite functions. The resulting annotated corpus is available upon request and can be accessed through a searchable online database.

A PropBank for Portuguese: the CINTIL-PropBank

António Branco, Catarina Carvalheiro, Sílvia Pereira, Sara Silveira, João Silva, Sérgio Castro and João Graça

With the CINTIL-International Corpus of Portuguese, an ongoing corpus annotated with fully flegded grammatical representation, sentences get not only a high level of lexical, morphological and syntactic annotation but also a semantic analysis that prepares the data to a manual specification step and thus opens the way for a number of tools and resources for which there is a great research focus at the present. This paper reports on the construction of a propbank that builds on CINTIL-DeepGramBank, with nearly 10 thousand sentences, on the basis of a deep linguistic grammar and on the process and the linguistic criteria guiding that construction, which makes possible to obtain a complete PropBank with both syntactic and semantic levels of linguistic annotation. Taking into account this and the promising scores presented in this study for inter-annotator agreement, CINTIL-PropBank presents itself as a great resource to train a semantic role labeller, one of our goals with this project.

Empty Argument Insertion in the Hindi PropBank

Ashwini Vaidya, Jinho D. Choi, Martha Palmer and Bhuvana Narasimhan

This paper examines both linguistic behavior and practical implication of empty argument insertion in the Hindi PropBank.

The Hindi PropBank is annotated on the Hindi Dependency Treebank, which contains some empty categories but not the empty arguments of verbs. In this paper, we analyze four kinds of empty arguments, *PRO*, *REL*, *GAP*, *pro*, and suggest effective ways of annotating these arguments. Empty arguments such as *PRO* and *REL* can be inserted deterministically; we present linguistically motivated rules that automatically insert these arguments with high accuracy. On the other hand, it is difficult to find deterministic rules to insert *GAP* and *pro*; for these arguments, we introduce a new annotation scheme that concurrently handles both semantic role labeling and empty category insertion, producing fast and high quality annotation. In addition, we present algorithms for finding antecedents of *REL* and *PRO*, and discuss why finding antecedents for some types of *PRO* is difficult.

Annotating Qualia Relations in Italian and French Complex Nominals

Pierrette Bouillon, Elisabetta Jezek, Chiara Melloni and Aurélie Picton

The goal of this paper is to provide an annotation scheme for compounds based on generative lexicon theory (GL, Pustejovsky, 1995; Bassac and Bouillon, 2001). This scheme has been tested on a set of compounds automatically extracted from the Europarl corpus (Koehn, 2005) both in Italian and French. The motivation is twofold. On the one hand, it should help refine existing compound classifications and better explain lexicalization in both languages. On the other hand, we hope that the extracted generalizations can be used in NLP, for example for improving MT systems or for query reformulation (Claveau, 2003). In this paper, we focus on the annotation scheme and its on going evaluation.

Semantic annotation of French corpora: animacy and verb semantic classes

Juliette Thuilier and Laurence Danlos

This paper presents a first corpus of French annotated for animacy and for verb semantic classes. The resource consists of 1,346 sentences extracted from three different corpora: the French Treebank (Abeillé and Barrier, 2004), the Est-Républicain corpus (CNRTL) and the ESTER corpus (ELRA). It is a set of parsed sentences, containing a verbal head subcategorizing two complements, with annotations on the verb and on both complements, in the TIGER XML format (Mengel and Lezius, 2000). The resource was manually annotated and manually corrected by three annotators. Animacy has been annotated following the categories of Zaenen et al. (2004). Measures of inter-annotator agreement are good (Multi-pi = 0.82 and Multi-kappa = 0.86 (k = 3, N = 2360)). As for verb semantic classes, we used three of the five levels of classification of an existing dictionary: 'Les Verbes du Français' (Dubois and Dubois-Charlier, 1997). For the higher level (generic classes), the measures of agreement are Multi-pi = 0.84 and Multi-kappa = 0.87 (k = 3, N = 1346). The inter-annotator agreements show that the annotated data are reliable for both animacy and verbal semantic classes.

Yes we can!? Annotating English modal verbs

Josef Ruppenhofer and Ines Rehbein

This paper presents an annotation scheme for English modal verbs together with sense-annotated data from the news domain. We describe our annotation scheme and discuss problematic cases for modality annotation based on the inter-annotator agreement during the annotation. Furthermore, we present experiments on automatic sense tagging, showing that our annotations do provide a valuable training resource for NLP systems.

An Annotation Scheme for Quantifier Scope Disambiguation

Mehdi Manshadi, James Allen and Mary Swift

Annotating natural language sentences with quantifier scoping has proved to be very hard. In order to overcome the challenge, previous work on building scope-annotated corpora has focused on sentences with two explicitly quantified noun phrases (NPs). Furthermore, it does not address the annotation of scopal operators or complex NPs such as plurals and definites. We present the first annotation scheme for quantifier scope disambiguation where there is no restriction on the type or the number of scope-bearing elements in the sentence. We discuss some of the most prominent complex scope phenomena encountered in annotating the corpus, such as plurality and type-token distinction, and present mechanisms to handle those phenomena.

Building Japanese Predicate-argument Structure Corpus using Lexical Conceptual Structure

Yuichiroh Matsubayashi, Yusuke Miyao and Akiko Aizawa

This paper introduces our study on creating a Japanese corpus that is annotated using semantically-motivated predicate-argument structures. We propose an annotation framework based on Lexical Conceptual Structure (LCS), where semantic roles of arguments are represented through a semantic structure decomposed by several primitive predicates. As a first stage of the project, we extended Jackendoff 's LCS theory to increase generality of expression and coverage for verbs frequently appearing in the corpus, and successfully created LCS structures for 60 frequent Japanese predicates in Kyoto university Text Corpus (KTC). In this paper, we report our framework for creating the corpus and the current status of creating an LCS dictionary for Japanese predicates.

Semantic Annotations in Japanese FrameNet: Comparing Frames in Japanese and English

Kyoko Ohara

Since 2008, the Japanese FrameNet (JFN, http://jfn.st.hc.keio.ac.jp/) project has been annotating the Balanced Corpus of Contemporary Written Japanese (BCCWJ), the first such corpus, officially released in October 2011. This paper reports annotation results of the book genre of BCCWJ (Ohara 2011, Ohara, Saito, Fujii & Sato 2011). Comparing the semantic frames needed to annotate BCCWJ with those that the FrameNet (FN) project (Fillmore and Baker 2009, Fillmore 2006) already has defined revealed that: 1) differences in the Japanese and English semantic frames often concern different perspectives and different lexical aspects exhibited by the two lexicons; and 2) in most of the cases where JFN defined new semantic frame for a word, the frame did not involve culture-specific scenes. We investigated the extent to which existing semantic frames originally defined for analyzing English words were used, annotating 810 sentences of the so-called core data of the book genre of BCCWJ. In the 810 sentences we were able to assign semantic frames to approximately 4000 words, although we could not assign any to 587 words. That is, of all the LUs in the sentences, we were able to identify semantic frames to about 87 per cent of them. In other words, the semantic frames already defined in FN for English could be used for 87 per cent of the Japanese LUs.

ConanDoyle-neg: Annotation of negation cues and their scope in Conan Doyle stories

Roser Morante and Walter Daelemans

In this paper we present ConanDoyle-neg, a corpus of stories by Conan Doyle annotated with negation information. The negation cues and their scope, as well as the event or property that is negated have been annotated by two annotators. The inter-annotator agreement is measured in terms of F-scores at scope level. It is higher for cues (94.88 and 92.77), less high for scopes (85.04 and 77.31), and lower for the negated event (79.23 and 80.67). The corpus is publicly available.

P16 - Document Classification, Text Categorisation

Thursday, May 24, 9:45

Chairperson: **Serge Sharoff** Poster Session

The Netlog Corpus. A Resource for the Study of Flemish Dutch Internet Language

Mike Kestemont, Claudia Peersman, Benny De Decker, Guy De Pauw, Kim Luyckx, Roser Morante, Frederik Vaassen, Janneke van de Loo and Walter Daelemans

Although in recent years numerous forms of Internet communication – such as e-mail, blogs, chat rooms and social network environments – have emerged, balanced corpora of Internet speech with trustworthy meta-information (e.g. age and gender) or linguistic annotations are still limited. In this paper we present a large corpus of Flemish Dutch chat posts that were collected from the Belgian online social network Netlog. For all of these posts we also acquired the users' profile information, making this corpus a unique resource for computational and sociolinguistic research. However, for analyzing such a corpus on a large scale, NLP tools are required for e.g. automatic POS tagging or lemmatization. Because many NLP tools fail to correctly analyze the surface forms of chat language usage, we propose to normalize this 'anomalous' input into a format suitable for existing NLP solutions for standard Dutch. Additionally, we have annotated a substantial part of the corpus (i.e. the Chatty subset) to provide a gold standard for the evaluation of future approaches to automatic (Flemish) chat language normalization.

Investigating Verbal Intelligence using the TF-IDF Approach

Kseniya Zablotskaya, Fernando Fernández Martínez and Wolfgang Minker

In this paper we investigated differences in language use of speakers yielding different verbal intelligence when they describe the same event. The work is based on a corpus containing descriptions of a short film and verbal intelligence scores of the speakers. For analyzing the monologues and the film transcript, the number of reused words, lemmas, n-grams, cosine similarity and other features were calculated and compared to each other for different verbal intelligence groups. The results showed that the similarity of monologues of higher verbal intelligence speakers was greater than of lower and average verbal intelligence participants. A possible explanation of this phenomenon is that candidates yielding higher verbal intelligence have a good short-term memory. In this paper we also checked a hypothesis that differences in vocabulary of speakers yielding different verbal intelligence are sufficient enough for good classification results. For proving this hypothesis, the Nearest Neighbor classifier was trained using TF-IDF vocabulary measures. The maximum achieved accuracy was 92.86%.

Diachronic Changes in Text Complexity in 20th Century English Language: An NLP Approach

Sanja Štajner and Ruslan Mitkov

A syntactically complex text may represent a problem for both comprehension by humans and various NLP tasks. A large number of studies in text simplification are concerned with this problem and their aim is to transform the given text into a simplified form in order to make it accessible to the wider audience. In this study, we were investigating what the natural tendency of texts is in 20th century English language. Are they becoming syntactically more complex over the years, requiring a higher literacy level and greater effort from the readers, or are they becoming simpler and easier to read? We examined several factors of text complexity (average sentence length, Automated Readability Index, sentence complexity and passive voice) in the 20th century for two main English language varieties - British and American, using the 'Brown family' of corpora. In British English, we compared the complexity of texts published in 1931, 1961 and 1991, while in American English we compared the complexity of texts published in 1961 and 1992. Furthermore, we demonstrated how the state-of-the-art NLP tools can be used for automatic extraction of some complex features from the raw text version of the corpora.

DeCour: a corpus of DEceptive statements in Italian COURts

Tommaso Fornaciari and Massimo Poesio

In criminal proceedings, sometimes it is not easy to evaluate the sincerity of oral testimonies. DECOUR - DEception in COURt corpus - has been built with the aim of training models suitable to discriminate, from a stylometric point of view, between sincere and deceptive statements. DECOUR is a collection of hearings held in four Italian Courts, in which the speakers lie in front of the judge. These hearings become the object of a specific criminal proceeding for calumny or false testimony, in which the deceptiveness of the statements of the defendant is ascertained. Thanks to the final Court judgment, that points out which lies are told, each utterance of the corpus has been annotated as true, uncertain or false, according to its degree of truthfulness. Since the judgment of deceptiveness follows a judicial inquiry, the annotation has been realized with a greater degree of confidence than ever before. Moreover, in Italy this is the first corpus of deceptive texts not relying on 'mock' lies created in laboratory conditions, but which has been collected in a natural environment.

French and German Corpora for Audience-based Text Type Classification

Amalia Todirascu, Sebastian Pado, Jennifer Krisch, Max Kisselew and Ulrich Heid

This paper presents some of the results of the CLASSYN project which investigated the classification of text according to audience-related text types. We describe the design principles and the properties of the French and German linguistically annotated corpora that we have created. We report on tools used to collect the data and on the quality of the syntactic annotation. The CLASSYN corpora comprise two text collections to investigate general text types difference between scientific and popular science text on the two domains of medical and computer science.

Irregularity Detection in Categorized Document Corpora

Borut Sluban, Senja Pollak, Roel Coesemans and Nada Lavrac

The paper presents an approach to extract irregularities in document corpora, where the documents originate from different sources and the analyst's interest is to find documents which are atypical for the given source. The main contribution of the paper is a voting-based approach to irregularity detection and its evaluation on a collection of newspaper articles from two sources: Western (UK and US) and local (Kenyan) media. The evaluation of a domain expert proves that the method is very effective in uncovering interesting irregularities in categorized document corpora.

Rhetorical Move Detection in English Abstracts: Multi-label Sentence Classifiers and their Annotated Corpora

Carmen Dayrell, Arnaldo Candido Jr., Gabriel Lima, Danilo Machado Jr., Ann Copestake, Valéria Feltrim, Stella Tagnin and Sandra Aluisio

The relevance of automatically identifying rhetorical moves in scientific texts has been widely acknowledged in the literature. This study focuses on abstracts of standard research papers written in English and aims to tackle a fundamental limitation of current machine-learning classifiers: they are mono-labeled, that is, a sentence can only be assigned one single label. However, such approach does not adequately reflect actual language use since a move can be realized by a clause, a sentence, or even several sentences. Here, we present MAZEA (Multi-label Argumentative Zoning for English Abstracts), a multi-label classifier which automatically identifies rhetorical moves in abstracts but allows for a given sentence to be assigned as many labels as appropriate.

We have resorted to various other NLP tools and used two large training corpora: (i) one corpus consists of 645 abstracts from physical sciences and engineering (PE) and (ii) the other corpus is made up of 690 from life and health sciences (LH). This paper presents our preliminary results and also discusses the various challenges involved in multi-label tagging and works towards satisfactory solutions. In addition, we also make our two training corpora publicly available so that they may serve as benchmark for this new task.

Unsupervised document zone identification using probabilistic graphical models

Andrea Varga, Daniel Preotiuc-Pietro and Fabio Ciravegna

Document zone identification aims to automatically classify sequences of text-spans (e.g. sentences) within a document into predefined zone categories. Current approaches to document zone identification mostly rely on supervised machine learning methods, which require a large amount of annotated data, which is often difficult and expensive to obtain. In order to overcome this bottleneck, we propose graphical models based on the popular Latent Dirichlet Allocation (LDA) model. The first model, which we call zoneLDA aims to cluster the sentences into zone classes using only unlabelled data. We also study an extension of zoneLDA called zoneLDAb, which makes distinction between common words and non-common words within the different zone types. We present results on two different domains: the scientific domain and the technical domain. For the latter one we propose a new document zone classification schema, which has been annotated over a collection of 689 documents, achieving a Kappa score of 85%. Overall our experiments show promising results for both of the domains, outperforming the baseline model. Furthermore, on the technical domain the performance of the models are comparable to the supervised approach using the same feature sets. We thus believe that graphical models are a promising avenue of research for automatic document zoning.

Improving K-Nearest Neighbor Efficacy for Farsi Text Classification

Mohammad Hossein Elahimanesh, Behrouz Minaei and Hossein Malekinezhad

One of the common processes in the field of text mining is text classification. Because of the complex nature of Farsi language, words with separate parts and combined verbs, the most of text classification systems are not applicable to Farsi texts. K-Nearest Neighbors (KNN) is one of the most popular used methods for text classification and presents good performance in experiments on different datasets. A method to improve the classification performance of KNN is proposed in this paper. Effects of removing or maintaining stop words, applying N-Grams with different lengths are also studied. For this study, a portion of a standard Farsi corpus called Hamshahri1 and articles of some archived newspapers are used. As the results indicate, classification efficiency improves by applying this approach especially when eight-grams indexing method and removing stop words are applied. Using N-grams with lengths more than 3 characters, presented very encouraging results for Farsi text classification. The Results of classification using our method are compared with the results obtained by mentioned related works.

P17 - Grammar and Syntax

Thursday, May 24, 9:45

Chairperson: **Eleni Efthimiou** Poster Session

Automatic Annotation and Manual Evaluation of the Diachronic German Corpus TüBa-D/DC

Erhard Hinrichs and Thomas Zastrow

This paper presents the Tübingen Baumbank des Deutschen Diachron (TüBa-D/DC), a linguistically annotated corpus of selected diachronic materials from the German Gutenberg Project. It was automatically annotated by a suite of NLP tools integrated into WebLicht, the linguistic chaining tool used in CLARIN-D. The annotation quality has been evaluated manually for a subcorpus ranging from Middle High German to Modern High German. The integration of the TüBa-D/DC into the CLARIN-D infrastructure includes metadata provision and harvesting as well as sustainable data storage in the Tübingen CLARIN-D center. The paper further provides an overview of the possibilities of accessing the TüBa-D/DC data. Methods for full-text search of the metadata and object data and for annotation-based search of the object data are described in detail. The WebLicht Service Oriented Architecture is used as an integrated environment for annotation based search of the TüBa-D/DC. WebLicht thus not only serves as the annotation platform for the TüBa-D/DC, but also as a generic user interface for accessing and visualizing it.

Grammatical Error Annotation for Korean Learners of Spoken English

Hongsuck Seo, Kyusong Lee, Gary Geunbae Lee, Soo-Ok Kweon and Hae-Ri Kim

The goal of our research is to build a grammatical error-tagged corpus for Korean learners of Spoken English dubbed Postech Learner Corpus. We collected raw story-telling speech from Korean university students. Transcription and annotation using the Cambridge Learner Corpus tagset were performed by six Korean

annotators fluent in English. For the annotation of the corpus, we developed an annotation tool and a validation tool. After comparing human annotation with machine-recommended error tags, unmatched errors were rechecked by a native annotator. We observed different characteristics between the spoken language corpus built in this study and an existing written language corpus.

Robust clause boundary identification for corpus annotation

Heiki-Jaan Kaalep and Kadri Muischnek

The paper describes a rule-based system for tagging clause boundaries, implemented for annotating the Estonian Reference Corpus of the University of Tartu, a collection of written texts containing ca 245 million running words and available for querying via Keeleveeb language portal. The system needs information about parts of speech and grammatical categories coded in the word-forms, i.e. it takes morphologically annotated text as input, but requires no information about the syntactic structure of the sentence. Among the strong points of our system we should mention identifying parenthesis and embedded clauses, i.e. clauses that are inserted into another clause dividing it into two separate parts in the linear text, for example a relative clause following its head noun. That enables a corpus query system to unite the otherwise divided clause, a feature that usually presupposes full parsing. The overall precision of the system is 95% and the recall is 96%. If “ordinary” clause boundary detection and parenthesis and embedded clause boundary detection are evaluated separately, then one can say that detecting an “ordinary” clause boundary (recall 98%, precision 96%) is an easier task than detecting an embedded clause (recall 79%, precision 100%).

A Corpus-based Study of the German Recipient Passive

Patrick Ziering, Sina Zarrieß and Jonas Kuhn

In this paper, we investigate the usage of a non-canonical German passive alternation for ditransitive verbs, the recipient passive, in naturally occuring corpus data. We propose a classifier that predicts the voice of a ditransitive verb based on the contextually determined properties its arguments. As the recipient passive is a low frequent phenomenon, we first create a special data set focussing on German ditransitive verbs which are frequently used in the recipient passive. We use a broad-coverage grammar-based parser, the German LFG parser, to automatically annotate our data set for the morpho-syntactic properties of the involved predicate arguments. We train a Maximum Entropy classifier on the automatically annotated sentences and achieve an accuracy of 98.05%, clearly outperforming the baseline that always predicts active voice baseline (94.6%).

Wordnet Based Lexicon Grammar for Polish

Zygmunt Vetulani

In the paper we present a long-term on-going project of a lexicon-grammar of Polish. It is based on our former research focusing mainly on morphological dictionaries, text understanding and related tools. By Lexicon Grammars we mean grammatical formalisms which are based on the idea that sentence is the fundamental unit of meaning and that grammatical information should be closely related to words. Organization of the grammatical knowledge into a lexicon results in a powerful NLP tool, particularly well suited to support heuristic parsing. The project is inspired by the achievements of Maurice Gross, Kazimierz Polanski and George Miller. We present the actual state of the project of a wordnet-like lexical network PolNet with particular emphasis on its verbal component, now being converted into the kernel of a lexicon grammar for Polish. We present various aspects of PolNet development and validation within the POLINT-112-SMS project. The reader is precisely informed on the current stage of the project.

A Galician Syntactic Corpus with Application to Intonation Modeling

Montserrat Arza, José M. García-Miguel, Francisco Campillo and Miguel Cuevas - Alonso

This paper will present the design of a Galician syntactic corpus with application to intonation modeling. A corpus of around 3000 sentences was designed with variation in the syntactic structure and the number of accent groups, and recorded by a professional speaker to study the influence on the prosodic structure.

A Search Tool for FrameNet Constructicon

Hiroaki Sato

The Berkeley FrameNet Project (BFN, https://framenet.icsi.berkeley.edu/fndrupal/) created descriptions of 73 “non-core” grammatical constructions, annotation of 50 of these constructions and about 1500 example sentences in its one year project “Beyond the Core: A Pilot Project on Cataloging Grammatical Constructions and Multiword Expressions in English” supported by the National Science Foundation. The project did not aim at building a full-fledged Construction Grammar, but the registry of English constructions created by this project, which is called Constructicon, provides a representative sample of the current coverage of English constructions (Lee-Goldman & Rhodes 2009). CxN Viewer is a search tool which I

have developed for Constructicon and the tool shows its typical English constructions on the web browser. CxN Viewer is a web application consisting of HTML files and JavaScript codes. The tool is a useful program that will benefit researchers working with the data annotated within the framework of BFN. CxN Viewer is a unicode-compliant application, and it can deal with constructions of other languages such as Spanish.

Annotating Errors in a Hungarian Learner Corpus

Markus Dickinson and Scott Ledbetter

We are developing and annotating a learner corpus of Hungarian, composed of student journals from three different proficiency levels written at Indiana University. Our annotation marks learner errors that are of different linguistic categories, including phonology, morphology, and syntax, but defining the annotation for an agglutinative language presents several issues. First, we must adapt an analysis that is centered on the morpheme rather than the word. Second, and more importantly, we see a need to distinguish errors from secondary corrections. We argue that although certain learner errors require a series of corrections to reach a target form, these secondary corrections, conditioned on those that come before, are our own adjustments that link the learner's productions to the target form and are not representative of the learner's internal grammar. In this paper, we report the annotation scheme and the principles that guide it, as well as examples illustrating its functionality and directions for expansion.

Text Simplification Tools for Spanish

Stefan Bott, Horacio Saggion and Simon Mille

In this paper we describe the development of a text simplification system for Spanish. Text simplification is the adaptation of a text to the special needs of certain groups of readers, such as language learners, people with cognitive difficulties and elderly people, among others. There is a clear need for simplified texts, but manual production and adaptation of existing texts is labour intensive and costly. Automatic simplification is a field which attracts growing attention in Natural Language Processing, but, to the best of our knowledge, there are no simplification tools for Spanish. We present a prototype for automatic simplification, which shows that the most important structural simplification operations can be successfully treated with an approach based on rules which can potentially be improved by statistical methods. For the development of this prototype we carried out a corpus study which aims at identifying the operations a text simplification system needs to carry out in order to produce an output similar to what human editors produce when they simplify texts.

CLIMB grammars: three projects using metagrammar engineering

Antske Fokkens, Tania Avgustinova and Yi Zhang

This paper introduces the CLIMB (Comparative Libraries of Implementations with Matrix Basis) methodology and grammars. The basic idea behind CLIMB is to use code generation as a general methodology for grammar development in order to create a more systematic approach to grammar development. The particular method used in this paper is closely related to the LinGO Grammar Matrix. Like the Grammar Matrix, resulting grammars are HPSG grammars that can map bidirectionally between strings and MRS representations. The main purpose of this paper is to provide insight into the process of using CLIMB for grammar development. In addition, we describe three projects that make use of this methodology or have concrete plans to adapt CLIMB in the future: CLIMB for Germanic languages, CLIMB for Slavic languages and CLIMB to combine two grammars of Mandarin Chinese. We present the first results that indicate feasibility and development time improvements for creating a medium to large coverage precision grammar.

An implementation of a Latvian resource grammar in Grammatical Framework

Peteris Paikens and Normunds Gruzitis

This paper describes an open-source Latvian resource grammar implemented in Grammatical Framework (GF), a programming language for multilingual grammar applications. GF differentiates between concrete grammars and abstract grammars: translation among concrete languages is provided via abstract syntax trees. Thus the same concrete grammar is effectively used for both language analysis and language generation. Furthermore, GF differentiates between general-purpose resource grammars and domain-specific application grammars that are built on top of the resource grammars. The GF resource grammar library (RGL) currently supports more than 20 languages that implement a common API. Latvian is the 13th official European Union language that is made available in the RGL. We briefly describe the grammatical features of Latvian and illustrate how they are handled in the multilingual framework of GF. We also illustrate some application areas of the Latvian resource grammar, and briefly discuss the limitations of the RGL and potential long-term improvements using frame semantics.

An Open Source Persian Computational Grammar

Shafqat Mumtaz Virk and Elnaz Abolahrar

Abstract In this paper, we describe a multilingual open-source computational grammar of Persian, developed in Grammatical Framework (GF) – A type-theoretical grammar formalism. We discuss in detail the structure of different syntactic (i.e. noun phrases, verb phrases, adjectival phrases, etc.) categories of Persian. First, we show how to structure and construct these categories individually. Then we describe how they are glued together to make well-formed sentences in Persian, while maintaining the grammatical features such as agreement, word order, etc. We also show how some of the distinctive features of Persian, such as the ezafe construction, are implemented in GF. In order to evaluate the grammar's correctness, and to demonstrate its usefulness, we have added support for Persian in a multilingual application grammar (the Tourist Phrasebook) using the reported resource grammar.

Reclassifying subcategorization frames for experimental analysis and stimulus generation

Paula Buttery and Andrew Caines

Researchers in the fields of psycholinguistics and neurolinguistics increasingly test their experimental hypotheses against probabilistic models of language. VALEX (Korhonen et al., 2006) is a large-scale verb lexicon that specifies verb usage as probability distributions over a set of 163 verb SUBCATEGORIZATION FRAMES (SCFs). VALEX has proved to be a popular computational linguistic resource and may also be used by psycho- and neurolinguists for experimental analysis and stimulus generation. However, a probabilistic model based upon a set of 163 SCFs often proves too fine grained for experimenters in these fields. Our goal is to simplify the classification by grouping the frames into genera–explainable clusters that may be used as experimental parameters. We adopted two methods for reclassification. One was a manual linguistic approach derived from verb argumentation and clause features; the other was an automatic, computational approach driven from a graphical representation of SCFs. The premise was not only to compare the results of two quite different methods for our own interest, but also to enable other researchers to choose whichever reclassification better suited their purpose (one being grounded purely in theoretical linguistics and the other in practical language engineering). The various classifications are available as an online resource to researchers.

Annotating progressive aspect constructions in the spoken section of the British National Corpus

Andrew Caines and Paula Buttery

We present a set of stand-off annotations for the ninety thousand sentences in the spoken section of the British National Corpus (BNC) which feature a progressive aspect verb group. These annotations may be matched to the original BNC text using the supplied document and sentence identifiers. The annotated features mostly relate to linguistic form: subject type, subject person and number, form of auxiliary verb, and clause type, tense and polarity. In addition, the sentences are classified for register, the formality of recording context: three levels of 'spontaneity' with genres such as sermons and scripted speech at the most formal level and casual conversation at the least formal. The resource has been designed so that it may easily be augmented with further stand-off annotations. Expert linguistic annotations of spoken data, such as these, are valuable for improving the performance of natural language processing tools in the spoken language domain and assist linguistic research in general.

P18 - Digital Libraries

Thursday, May 24, 9:45

Chairperson: **Monica Monachini** Poster Session

BUCEADOR, a multi-language search engine for digital libraries

Jordi Adell, Antonio Bonafonte, Antonio Cardenal, Marta R. Costa-Jussà, José A. R. Fonollosa, Asunción Moreno, Eva Navas and Eduardo R. Banga

This paper presents a web-based multimedia search engine built within the Buceador (www.buceador.org) research project. A proof-of-concept tool has been implemented which is able to retrieve information from a digital library made of multimedia documents in the 4 official languages in Spain (Spanish, Basque, Catalan and Galician). The retrieved documents are presented in the user language after translation and dubbing (the four previous languages + English). The paper presents the tool functionality, the architecture, the digital library and provide some information about the technology involved in the fields of automatic speech recognition, statistical machine translation, text-to-speech synthesis and information retrieval. Each technology has been adapted to the purposes of the presented tool as well as to interact with the rest of the technologies involved.

A tool for enhanced search of multilingual digital libraries of e-journals

Ranka Stanković, Cvetana Krstev, Ivan Obradović, Aleksandra Trtovac and Miloš Utvić

This paper outlines the main features of Bibliša, a tool that

offers various possibilities of enhancing queries submitted to large collections of TMX documents generated from aligned parallel articles residing in multilingual digital libraries of e-journals. The queries initiated by a simple or multiword keyword, in Serbian or English, can be expanded by Bibliša, both semantically and morphologically, using different supporting monolingual and multilingual resources, such as wordnets and electronic dictionaries. The tool operates within a complex system composed of several modules including a web application, which makes it readily accessible on the web. Its functionality has been tested on a collection of 44 TMX documents generated from articles published bilingually by the journal INFOtecha, yielding encouraging results. Further enhancements of the tool are underway, with the aim of transforming it from a powerful full-text and metadata search tool, to a useful translator's aid, which could be of assistance both in reviewing terminology used in context and in refining the multilingual resources used within the system.

A Graphical Citation Browser for the ACL Anthology

Benjamin Weitz and Ulrich Schäfer

Navigation in large scholarly paper collections is tedious and not well supported in most scientific digital libraries. We describe a novel browser-based graphical tool implemented using HTML5 Canvas. It displays citation information extracted from the paper text to support useful navigation. The tool is implemented using a client/server architecture. A citation graph of the digital library is built in the memory of the server. On the client side, egdes of the displayed citation (sub)graph surrounding a document are labeled with keywords signifying the kind of citation made from one document to another. These keywords were extracted using NLP tools such as tokenizer, sentence boundary detection and part-of-speech tagging applied to the text extracted from the original PDF papers (currently 22,500). By clicking on an egde, the user can inspect the corresponding citation sentence in context, in most cases even also highlighted in the original PDF layout. The system is publicly accessible as part of the ACL Anthology Searchbench.

LDC Language Resource Database: Building a Bibliographic Database

Eleftheria Ahtaridis, Christopher Cieri and Denise DiPersio

The Linguistic Data Consortium (LDC) creates and provides language resources (LRs) including data, tools and specifications. In order to assess the impact of these LRs and to support both LR users and authors, LDC is collecting metadata about and URLs for research papers that introduce, describe, critique, extend or rely upon LDC LRs. Current collection efforts focus on papers published in journals and conference proceedings that are available online. To date, nearly 300, or over half of the LRs LDC distributes have been searched for extensively and almost 8000 research papers about these LRs have been documented. This paper discusses the issues with collecting references and includes preliminary analysis of those results. The remaining goals of the project are also outlined.

Matching Cultural Heritage items to Wikipedia

Eneko Agirre, Ander Barrena, Oier Lopez de Lacalle, Aitor Soroa, Samuel Fernando and Mark Stevenson

Digitised Cultural Heritage (CH) items usually have short descriptions and lack rich contextual information. Wikipedia articles, on the contrary, include in-depth descriptions and links to related articles, which motivate the enrichment of CH items with information from Wikipedia. In this paper we explore the feasibility of finding matching articles in Wikipedia for a given Cultural Heritage item. We manually annotated a random sample of items from Europeana, and performed a qualitative and quantitative study of the issues and problems that arise, showing that each kind of CH item is different and needs a nuanced definition of what "matching article" means. In addition, we test a well-known wikification (aka entity linking) algorithm on the task. Our results indicate that a substantial number of items can be effectively linked to their corresponding Wikipedia article.

O21 - Speech Corpora and Tools

Thursday, May 24, 11:45

Chairperson: **Robrecht Comeyne** Oral Session

Creating a Data Collection for Evaluating Rich Speech Retrieval

Maria Eskevich, Gareth J.F. Jones, Martha Larson and Roeland Ordelman

We describe the development of a test collection for the investigation of speech retrieval beyond identification of relevant content. This collection focuses on satisfying user information needs for queries associated with specific types of speech acts. The collection is based on an archive of the Internet video from Internet video sharing platform (blip.tv), and was provided by the MediaEval benchmarking initiative. A crowdsourcing approach was used to identify segments in the video data which contain speech acts, to create a description of the video containing the act and to generate search queries designed to refind this speech act. We describe and reflect on our experiences with crowdsourcing

this test collection using the Amazon Mechanical Turk platform. We highlight the challenges of constructing this dataset, including the selection of the data source, design of the crowdsouring task and the specification of queries and relevant items.

The Political Speech Corpus of Bulgarian

Petya Osenova and Kiril Simov

The paper introduces the Political Speech Corpus of Bulgarian. First, its current state has been discussed with respect to its size, coverage, genre specification and related online services. Then, the focus goes to the annotation details. On the one hand, the layers of linguistic annotation are presented. On the other hand, the compatibility with CLARIN technical Infrastructure is explained. Also, some user-based scenarios are mentioned to demonstrate the corpus services and applicability.

SPPAS: a tool for the phonetic segmentation of speech

Brigitte Bigi

SPPAS is a tool to produce automatic annotations which include utterance, word, syllabic and phonemic segmentations from a recorded speech sound and its transcription. SPPAS is distributed under the terms of the GNU Public License. It was successfully applied during the Evalita 2011 campaign, on Italian map-task dialogues. It can also deal with French, English and Chinese and there is an easy way to add other languages. The paper describes the development of resources and free tools, consisting of acoustic models, phonetic dictionaries, and libraries and programs to deal with these data. All of them are publicly available.

Orthographic Transcription: which enrichment is required for phonetization?

Brigitte Bigi, Pauline Péri and Roxane Bertrand

This paper addresses the problem of the enrichment of transcriptions in the perspective of an automatic phonetization. Phonetization is the process of representing sounds with phonetic signs. There are two general ways to construct a phonetization process: rule based systems (with rules based on inference approaches or proposed by expert linguists) and dictionary based solutions which consist in storing a maximum of phonological knowledge in a lexicon. In both cases, phonetization is based on a manual transcription. Such a transcription is established on the basis of conventions that can differ depending on their working out context. This present study focuses on three different enrichments of such a transcription. Evaluations compare phonetizations obtained from automatic systems to a reference phonetized manually. The test corpus is made of three types of speech: conversational speech, read speech and political debate. A specific algorithm for the rule-based system is proposed to deal with enrichments. The final system obtained a phonetization of about 95.2% correct (from 3.7% to 5.6% error rates depending on the corpus).

O22 - Machine Translation and Evaluation (2)

Thursday, May 24, 11:45

Chairperson: **François Yvon** Oral Session

Error profiling for evaluation of machine-translated text: a Polish-English case study

Sandra Weiss and Lars Ahrenberg

We present a study of Polish-English machine translation, where the impact of various types of errors on cohesion and comprehensibility of the translations were investigated. The following phenomena are in focus: (i) The most common errors produced by current state-of-the-art MT systems for Polish-English MT. (ii) The effect of different types of errors on text cohesion. (iii) The effect of different types of errors on readers' understanding of the translation. We found that errors of incorrect and missing translations are the most common for current systems, while the category of non-translated words had the most negative impact on comprehension. All three of these categories contributed to the breaking of cohesive chains. The correlation between number of errors found in a translation and number of wrong answers in the comprehension tests was low. Another result was that non-native speakers of English performed at least as good as native speakers on the comprehension tests.

Two Phase Evaluation for Selecting Machine Translation Services

Chunqi Shi, Donghui Lin, Masahiko Shimada and Toru Ishida

An increased number of machine translation services are now available. Unfortunately, none of them can provide adequate translation quality for all input sources. This forces the user to select from among the services according to his needs. However, it is tedious and time consuming to perform this manual selection. Our solution, proposed here, is an automatic mechanism that can select the most appropriate machine translation service. Although evaluation methods are available, such as BLEU, NIST, WER, etc., their evaluation results are not unanimous regardless of the translation sources. We proposed a two-phase architecture for selecting translation services. The first phase uses a data-driven classification to allow the most appropriate evaluation method to be selected according to each translation source. The second

phase selects the most appropriate machine translation result by the selected evaluation method. We describe the architecture, detail the algorithm, and construct a prototype. Tests show that the proposal yields better translation quality than employing just one machine translation service.

Italian and Spanish Null Subjects. A Case Study Evaluation in an MT Perspective.

Lorenza Russo, Sharid Loáiciga and Asheesh Gulati

Thanks to their rich morphology, Italian and Spanish allow pro-drop pronouns, i.e., non lexically-realized subject pronouns. Here we distinguish between two different types of null subjects: personal pro-drop and impersonal pro-drop. We evaluate the translation of these two categories into French, a non pro-drop language, using Its-2, a transfer-based system developed at our laboratory; and Moses, a statistical system. Three different corpora are used: two subsets of the Europarl corpus and a third corpus built using newspaper articles. Null subjects turn out to be quantitatively important in all three corpora, but their distribution varies depending on the language and the text genre though. From a MT perspective, translation results are determined by the type of pro-drop and the pair of languages involved. Impersonal pro-drop is harder to translate than personal pro-drop, especially for the translation from Italian into French, and a significant portion of incorrect translations consists of missing pronouns.

On the practice of error analysis for machine translation evaluation

Sara Stymne and Lars Ahrenberg

Error analysis is a means to assess machine translation output in qualitative terms, which can be used as a basis for the generation of error profiles for different systems. As for other subjective approaches to evaluation it runs the risk of low inter-annotator agreement, but very often in papers applying error analysis to MT, this aspect is not even discussed. In this paper, we report results from a comparative evaluation of two systems where agreement initially was low, and discuss the different ways we used to improve it. We compared the effects of using more or less fine-grained taxonomies, and the possibility to restrict analysis to short sentences only. We report results on inter-annotator agreement before and after measures were taken, on error categories that are most likely to be confused, and on the possibility to establish error profiles also in the absence of a high inter-annotator agreement.

O23 - Semantic Resources

Thursday, May 24, 11:45

Chairperson: **Zygmunt Vetulani** Oral Session

Identifying equivalents of specialized verbs in a bilingual comparable corpus of judgments: A frame-based methodology

Janine Pimentel

Multilingual terminological resources do not always include the equivalents of specialized verbs that occur in legal texts. This study aims to bridge that gap by proposing a methodology to assign the equivalents of this kind of predicative units. We use a comparable corpus of judgments produced by the Supreme Court of Canada and by the Supremo Tribunal de Justiça de Portugal. From this corpus, 200 English and Portuguese verbs are selected. The description of the verbs is based on the theory of Frame Semantics (Fillmore 1977, 1977, 1982, 1985) as well as on the FrameNet methodology (Ruppenhofer et al. 2010). Specialized verbs are said to evoke a semantic frame, a sort of conceptual scenario in which a number of mandatory elements play specific roles (e.g. the role of judge, the role of defendant). Given that semantic frames are language independent to a fair degree (Boas 2005; Baker 2009), the labels attributed to each of the 76 identified frames (e.g. [Crime], [Regulations]) were used to group together 165 pairs of candidate equivalents. 71% of them are full equivalents, whereas 29% are only partial equivalents.

Logical metonymies and qualia structures: an annotated database of logical metonymies for German

Alessandra Zarcone and Stefan Rued

Logical metonymies like“The author began the book”involve the interpretation of events that are not realized in the sentence (Covert events: ->“writing the book"). The Generative Lexicon (Pustejovsky 1995) provides a qualia-based account of covert event interpretation, claiming that the covert event is retrieved from the qualia structure of the object. Such a theory poses the question of to what extent covert events in logical metonymies can be accounted for by qualia structures. Building on previous work on English, we present a corpus study for German verbs ("anfangen (mit)",“aufhoeren (mit)",“beenden",“beginnen (mit)",“geniessen", based on data obtained from the deWaC corpus. We built a corpus of logical metonymies, which were manually annotated and compared with the qualia structures of their objects, then we contrasted annotation results from two expert annotators for metonymies ("The author began the book") and long forms ("The author began reading the book") across

verbs. Our annotation was evaluated on a sample of sentences annotated by a group of naive annotators on a crowdsourcing platform. The logical metonymy database (2661 metonymies and 1886 long forms) with two expert annotations is freely available for scientific research purposes.

Modality in Text: a Proposal for Corpus Annotation

Iris Hendrickx, Amália Mendes and Silvia Mencarelli

We present a annotation scheme for modality in Portuguese. In our annotation scheme we have tried to combine a more theoretical linguistic viewpoint with a practical annotation scheme that will also be useful for NLP research but is not geared towards one specific application. Our notion of modality focuses on the attitude and opinion of the speaker, or of the subject of the sentence. We validated the annotation scheme on a corpus sample of approximately 2000 sentences that we fully annotated with modal information using the MMAX2 annotation tool to produce XML annotation. We discuss our main findings and give attention to the difficult cases that we encountered as they illustrate the complexity of modality and its interactions with other elements in the text.

DBpedia: A Multilingual Cross-domain Knowledge Base

Pablo Mendes, Max Jakob and Christian Bizer

The DBpedia project extracts structured information from Wikipedia editions in 97 different languages and combines this information into a large multi-lingual knowledge base covering many specific domains and general world knowledge. The knowledge base contains textual descriptions (titles and abstracts) of concepts in up to 97 languages. It also contains structured knowledge that has been extracted from the infobox systems of Wikipedias in 15 different languages and is mapped onto a single consistent ontology by a community effort. The knowledge base can be queried using the SPARQL query language and all its data sets are freely available for download. In this paper, we describe the general DBpedia knowledge base and as well as the DBpedia data sets that specifically aim at supporting computational linguistics tasks. These task include Entity Linking, Word Sense Disambiguation, Question Answering, Slot Filling and Relationship Extraction. These use cases are outlined, pointing at added value that the structured data of DBpedia provides.

O24 - Trends in Corpora

Thursday, May 24, 11:45

Chairperson: **Victoria Arranz** Oral Session

A corpus of general and specific sentences from news

Annie Louis and Ani Nenkova

We present a corpus of sentences from news articles that are annotated as general or specific. We employed annotators on Amazon Mechanical Turk to mark sentences from three kinds of news articles—reports on events, finance news and science journalism. We introduce the resulting corpus, with focus on annotator agreement, proportion of general/specific sentences in the articles and results for automatic classification of the two sentence types.

Brand Pitt: A Corpus to Explore the Art of Naming

Gozde Ozbal, Carlo Strapparava and Marco Guerini

The name of a company or a brand is the key element to a successful business. A good name is able to state the area of competition and communicate the promise given to customers by evoking semantic associations. Although various resources provide distinct tips for inventing creative names, little research was carried out to investigate the linguistic aspects behind the naming mechanism. Besides, there might be latent methods that copywriters unconsciously use. In this paper, we describe the annotation task that we have conducted on a dataset of creative names collected from various resources to create a gold standard for linguistic creativity in naming. Based on the annotations, we compile common and latent methods of naming and explore the correlations among linguistic devices, provoked effects and business domains. This resource represents a starting point for a corpus based approach to explore the art of naming.

The WeSearch Corpus, Treebank, and Treecache – A Comprehensive Sample of User-Generated Content

Jonathon Read, Dan Flickinger, Rebecca Dridan, Stephan Oepen and Lilja Øvrelid

We present the WeSearch Data Collection (WDC)—a freely redistributable, partly annotated, comprehensive sample of User-Generated Content. The WDC contains data extracted from a range of genres of varying formality (user forums, product review sites, blogs and Wikipedia) and covers two different domains (NLP and Linux). In this article, we describe the data selection and extraction process, with a focus on the extraction of linguistic

content from different sources. We present the format of syntacto-semantic annotations found in this resource and present initial parsing results for these data, as well as some reflections following a first round of treebanking.

Collecting humorous expressions from a community-based question-answering-service corpus

Masashi Inoue and Toshiki Akagi

We proposed a method of collecting humorous expressions from an online community-based question-answering (CQA) corpus where some users post a variety of questions and other users post relevant answers. Although the service is created for the purpose of knowledge exchange, there are users who enjoy posting humorous responses. Therefore, the corpus contains many interesting humour communication examples that might be useful in understanding the nature of online communications and variations in humour. Considering the size of 3; 116; 009 topics, it is necessary to introduce automation in the collection process. However, due to the context dependency of humour expressions, it is hard to collect them automatically by using keywords or key phrases. Our method uses natural language processing based on dissimilarity criteria between answer texts. By using this method, we can collect humour expressions more efficiently than by manual exploration: 30 times more examples per hour.

P19 - Treebanks

Thursday, May 24, 11:45

Chairperson: **Menno van Zaanen** Poster Session

Further Developments in Treebank Error Detection Using Derivation Trees

Seth Kulick, Ann Bies and Justin Mott

This work describes how derivation tree fragments based on a variant of Tree Adjoining Grammar (TAG) can be used to check treebank consistency. Annotation of word sequences are compared both for their internal structural consistency, and their external relation to the rest of the tree. We expand on earlier work in this area in three ways. First, we provide a more complete description of the system, showing how a naive use of TAG structures will not work, leading to a necessary refinement. We also provide a more complete account of the processing pipeline, including the grouping together of structurally similar errors and their elimination of duplicates. Second, we include the new experimental external relation check to find an additional class of errors. Third, we broaden the evaluation to include both the internal and external relation checks, and evaluate the system on both an Arabic and English treebank. The evaluation has been successful enough that the internal check has been integrated into the standard pipeline for current English treebank construction at the Linguistic Data Consortium

Parallel Aligned Treebanks at LDC: New Challenges Interfacing Existing Infrastructures

Xuansong Li, Stephanie Strassel, Stephen Grimes, Safa Ismael, Mohamed Maamouri, Ann Bies and Nianwen Xue

Parallel aligned treebanks (PAT) are linguistic corpora annotated with morphological and syntactic structures that are aligned at sentence as well as sub-sentence levels. They are valuable resources for improving machine translation (MT) quality. Recently, there has been an increasing demand for such data, especially for divergent language pairs. The Linguistic Data Consortium (LDC) and its academic partners have been developing Arabic-English and Chinese-English PATs for several years. This paper describes the PAT corpus creation effort for the program GALE (Global Autonomous Language Exploitation) and introduces the potential issues of scaling up this PAT effort for the program BOLT (Broad Operational Language Translation). Based on existing infrastructures and in the light of current annotation process, challenges and approaches, we are exploring new methodologies to address emerging challenges in constructing PATs, including data volume bottlenecks, dialect issues of Arabic languages, and new genre features related to rapidly changing social media. Preliminary experimental results are presented to show the feasibility of the approaches proposed.

Expanding Arabic Treebank to Speech: Results from Broadcast News

Mohamed Maamouri, Ann Bies and Seth Kulick

Treebanking a large corpus of relatively structured speech transcribed from various Arabic Broadcast News (BN) sources has allowed us to begin to address the many challenges of annotating and parsing a speech corpus in Arabic. The now completed Arabic Treebank BN corpus consists of 432,976 source tokens (517,080 tree tokens) in 120 files of manually transcribed news broadcasts. Because news broadcasts are predominantly scripted, most of the transcribed speech is in Modern Standard Arabic (MSA). As such, the lexical and syntactic structures are very similar to the MSA in written newswire data. However, because this is spoken news, cross-linguistic speech effects such as restarts, fillers, hesitations, and repetitions are common. There is also a certain amount of dialect data present in the BN corpus, from on-the-street interviews and similar informal contexts. In this paper, we describe the finished corpus and focus on some of the necessary additions to our annotation guidelines, along with some

of the technical challenges of a treebanked speech corpus and an initial parsing evaluation for this data. This corpus will be available to the community in 2012 as an LDC publication.

Propbank-Br: a Brazilian Treebank annotated with semantic role labels

Magali Sanches Duran and Sandra Maria Aluísio

This paper reports the annotation of a Brazilian Portuguese Treebank with semantic role labels following Propbank guidelines. A different language and a different parser output impact the task and require some decisions on how to annotate the corpus. Therefore, a new annotation guide – called Propbank-Br - has been generated to deal with specific language phenomena and parser problems. In this phase of the project, the corpus was annotated by a unique linguist. The annotation task reported here is inserted in a larger projet for the Brazilian Portuguese language. This project aims to build Brazilian verbs frames files and a broader and distributed annotation of semantic role labels in Brazilian Portuguese, allowing inter-annotator agreement measures. The corpus, available in web, is already being used to build a semantic tagger for Portuguese language.

Joint Grammar and Treebank Development for Mandarin Chinese with HPSG

Yi Zhang, Rui Wang and Yu Chen

We present the ongoing development of MCG, a linguistically deep and precise grammar for Mandarin Chinese together with its accompanying treebank, both based on the linguistic framework of HPSG, and using MRS as the semantic representation. We highlight some key features of our grammar design, and review a number of challenging phenomena, with comparisons to alternative linguistic treatments and implementations. One of the distinguishing characteristics of our approach is the tight integration of grammar and treebank development. The two-step treebank annotation procedure benefits from the efficiency of the discriminant-based annotation approach, while giving the annotators full freedom of producing extra-grammatical structures. This not only allows the creation of a precise and full-coverage treebank with an imperfect grammar, but also provides prompt feedback for grammarians to identify the errors in the grammar design and implementation. Preliminary evaluation and error analysis shows that the grammar already covers most of the core phenomena for Mandarin Chinese, and the treebank annotation procedure reaches a stable speed of 35 sentences per hour with satisfying quality.

A tree is a Baum is an árbol is a sach'a: Creating a trilingual treebank

Annette Rios and Anne Göhring

This paper describes the process of constructing a trilingual parallel treebank. While for two of the involved languages, Spanish and German, there are already corpora with well-established annotation schemes available, this is not the case with the third language: Cuzco Quechua (ISO 639-3:quz), a low-resourced, non-standardized language for which we had to define a linguistically plausible annotation scheme first.

German and English Treebanks and Lexica for Tree-Adjoining Grammars

Miriam Kaeshammer and Vera Demberg

We present a treebank and lexicon for German and English, which have been developed for PLTAG parsing. PLTAG is a psycholinguistically motivated, incremental version of tree-adjoining grammar (TAG). The resources are however also applicable to parsing with other variants of TAG. The German PLTAG resources are based on the TIGER corpus and, to the best of our knowledge, constitute the first scalable German TAG grammar. The English PLTAG resources go beyond existing resources in that they include the NP annotation by (Vadas and Curran, 2007), and include the prediction lexicon necessary for PLTAG.

Prague Dependency Style Treebank for Tamil

Loganathan Ramasamy and Zdeněk Žabokrtský

Annotated corpora such as treebanks are important for the development of parsers, language applications as well as understanding of the language itself. Only very few languages possess these scarce resources. In this paper, we describe our efforts in syntactically annotating a small corpora (600 sentences) of Tamil language. Our annotation is similar to Prague Dependency Treebank (PDT) and consists of annotation at 2 levels or layers: (i) morphological layer (m-layer) and (ii) analytical layer (a-layer). For both the layers, we introduce annotation schemes i.e. positional tagging for m-layer and dependency relations for a-layers. Finally, we discuss some of the issues in treebank development for Tamil.

Treebanking by Sentence and Tree Transformation: Building a Treebank to support Question Answering in Portuguese

Patricia Gonçalves, Rita Santos and António Branco

This paper presents CINTIL-QATreebank, a treebank composed of Portuguese sentences that can be used to support the

development of Question Answering systems. To create this treebank, we use declarative sentences from the pre-existing CINTIL-Treebank and manually transform their syntactic structure into a non-declarative sentence. Our corpus includes two clause types: interrogative and imperative clauses. CINTIL-QATreebank can be used in language science and techology general research, but it was developed particularly for the development of automatic Question Answering systems. The non-declarative entences are annotated with several layers of linguistic information, namely (i) trees with information on constituency and grammatical function; (ii) sentence type; (iii) interrogative pronoun; (iv) question type; and (v) semantic type of expected answer. Moreover, these non-declarative sentences are paired with their declarative counterparts and associated with the expected answer snippets.

Croatian Dependency Treebank: Recent Development and Initial Experiments

Dasa Berovic, Zeljko Agic and Marko Tadić

We present the current state of development of the Croatian Dependency Treebank – with special empahsis on adapting the Prague Dependency Treebank formalism to Croatian language specifics – and illustrate its possible applications in an experiment with dependency parsing using MaltParser. The treebank currently contains approximately 2870 sentences, out of which the 2699 sentences and 66930 tokens were used in this experiment. Three linear-time projective algorithms implemented by the MaltParser system – Nivre eager, Nivre standard and stack projective – running on default settings were used in the experiment. The highest performing system, implementing the Nivre eager algorithm, scored (LAS 71.31 UAS 80.93 LA 83.87) within our experiment setup. The results obtained serve as an illustration of treebank's usefulness in natural language processing research and as a baseline for further research in dependency parsing of Croatian.

A GUI to Detect and Correct Errors in Hindi Dependency Treebank

Rahul Agarwal, Bharat Ram Ambati and Anil Kumar Singh

A treebank is an important resource for developing many NLP based tools. Errors in the treebank may lead to error in the tools that use it. It is essential to ensure the quality of a treebank before it can be deployed for other purposes. Automatic (or semi-automatic) detection of errors in the treebank can reduce the manual work required to find and remove errors. Usually, the errors found automatically are manually corrected by the annotators. There is not much work reported so far on error correction tools which helps the annotators in correcting errors efficiently. In this paper, we present such an error correction tool that is an extension of the error detection method described earlier (Ambati et al., 2010; Ambati et al., 2011; Agarwal et al., 2012).

From Grammar Rule Extraction to Treebanking: A Bootstrapping Approach

Masood Ghayoomi

Most of the reliable language resources are developed via human supervision. Developing supervised annotated data is hard and tedious, and it will be very time consuming when it is done totally manually; as a result, various types of annotated data, including treebanks, are not available for many languages. Considering that a portion of the language is regular, we can define regular expressions as grammar rules to recognize the strings which match the regular expressions, and reduce the human effort to annotate further unseen data. In this paper, we propose an incremental bootstrapping approach via extracting grammar rules when no treebank is available in the first step. Since Persian suffers from lack of available data sources, we have applied our method to develop a treebank for this language. Our experiment shows that this approach significantly decreases the amount of manual effort in the annotation process while enlarging the treebank.

The IULA Treebank

Montserrat Marimon, Beatríz Fisas, Núria Bel, Marta Villegas, Jorge Vivaldi, Sergi Torner, Mercè Lorente, Silvia Vázquez and Marta Villegas

This paper describes on-going work for the construction of a new treebank for Spanish, The IULA Treebank. This new resource will contain about 60,000 richly annotated sentences as an extension of the already existing IULA Technical Corpus which is only PoS tagged. In this paper we have focused on describing the work done for defining the annotation process and the treebank design principles. We report on how the used framework, the DELPH-IN processing framework, has been crucial in the design principles and in the bootstrapping strategy followed, especially in what refers to the use of stochastic modules for reducing parsing overgeneration. We also report on the different evaluation experiments carried out to guarantee the quality of the already available results.

Specifying Treebanks, Outsourcing Parsebanks: FinnTreeBank 3

Atro Voutilainen, Kristiina Muhonen, Tanja Purtonen and Krister Lindén

Corpus-based treebank annotation is known to result in incomplete coverage of mid- and low-frequency linguistic constructions: the linguistic representation and corpus annotation quality are sometimes suboptimal. Large descriptive grammars cover also many mid- and low-frequency constructions. We argue for use of large descriptive grammars and their sample sentences as a basis for specifying higher-coverage grammatical representations. We present an sample case from an ongoing project (FIN-CLARIN FinnTreeBank) where an grammatical representation is documented as an annotator's manual alongside manual annotation of sample sentences extracted from a large descriptive grammar of Finnish. We outline the linguistic representation (morphology and dependency syntax) for Finnish, and show how the resulting 'Grammar Definition Corpus' and the documentation is used as a task specification for an external subcontractor for building a parser engine for use in morphological and dependency syntactic analysis of large volumes of Finnish for parsebanking purposes. The resulting corpus, FinnTreeBank 3, is due for release in June 2012, and will contain tens of millions of words from publicly available corpora of Finnish with automatic morphological and dependency syntactic analysis, for use in research on the corpus linguistics and language engineering.

The Parallel-TUT: a multilingual and multiformat treebank

Cristina Bosco, Manuela Sanguinetti and Leonardo Lesmo

The paper introduces an ongoing project for the development of a parallel treebank for Italian, English and French, i.e. Parallel–TUT, or simply ParTUT. For the development of this resource, both the dependency and constituency-based formats of the Italian Turin University Treebank (TUT) have been applied to a preliminary dataset, which includes the whole text of the Universal Declaration of Human Rights, and sentences from the JRC-Acquis Multilingual Parallel Corpus and the Creative Commons licence. The focus of the project is mainly on the quality of the annotation and the investigation of some issues related to the alignment of data that can be allowed by the TUT formats, also taking into account the availability of conversion tools for display data in standard ways, such as Tiger–XML and CoNLL formats. It is, in fact, our belief that increasing the portability of our treebank could give us the opportunity to access resources and tools provided by other research groups, especially at this stage of the project, where no particular tool – compatible with the TUT format – is available in order to tackle the alignment problems.

Irish Treebanking and Parsing: A Preliminary Evaluation

Teresa Lynn, Ozlem Cetinoglu, Jennifer Foster, Elaine Uí Dhonnchadha, Mark Dras and Josef van Genabith

Language resources are essential for linguistic research and the development of NLP applications. Low-density languages, such as Irish, therefore lack significant research in this area. This paper describes the early stages in the development of new language resources for Irish – namely the first Irish dependency treebank and the first Irish statistical dependency parser. We present the methodology behind building our new treebank and the steps we take to leverage upon the few existing resources. We discuss language-specific choices made when defining our dependency labelling scheme, and describe interesting Irish language characteristics such as prepositional attachment, copula, and clefting. We manually develop a small treebank of 300 sentences based on an existing POS-tagged corpus and report an inter-annotator agreement of 0.7902. We train MaltParser to achieve preliminary parsing results for Irish and describe a bootstrapping approach for further stages of development.

P20 - Parsing

Thursday, May 24, 11:45

Chairperson: **Antonio Moreno-Sandoval** Poster Session

Automatic Extraction and Evaluation of Arabic LFG Resources

Mohammed Attia, Khaled Shaalan, Lamia Tounsi and Josef van Genabith

This paper presents the results of an approach to automatically acquire large-scale, probabilistic Lexical-Functional Grammar (LFG) resources for Arabic from the Penn Arabic Treebank (ATB). Our starting point is the earlier, work of (Tounsi et al., 2009) on automatic LFG f(eature)-structure annotation for Arabic using the ATB. They exploit tree configuration, POS categories, functional tags, local heads and trace information to annotate nodes with LFG feature-structure equations. We utilize this annotation to automatically acquire grammatical function (dependency) based subcategorization frames and paths linking long-distance dependencies (LDDs). Many state-of-the-art treebank-based probabilistic parsing approaches are scalable and robust but often also shallow: they do not capture LDDs and represent only local information. Subcategorization frames

and LDD paths can be used to recover LDDs from such parser output to capture deep linguistic information. Automatic acquisition of language resources from existing treebanks saves time and effort involved in creating such resources by hand. Moreover, data-driven automatic acquisition naturally associates probabilistic information with subcategorization frames and LDD paths. Finally, based on the statistical distribution of LDD path types, we propose empirical bounds on traditional regular expression based functional uncertainty equations used to handle LDDs in LFG.

Rule-Based Detection of Clausal Coordinate Ellipsis

Kristiina Muhonen and Tanja Purtonen

With our experiment, we show how we can detect and annotate clausal coordinate ellipsis with Constraint Grammar rules. We focus on such an elliptical structure in which there are two coordinated clauses, and the latter one lacks a verb. For example, the sentence This belongs to me and that to you demonstrates the ellipsis in question, namely gapping. The Constraint Grammar rules are made for a Finnish parsebank, FinnTreeBank. The FinnTreeBank project is building a parsebank in the dependency syntactic framework in which verbs are central since other sentence elements depend on them. Without correct detection of omitted verbs, the syntactic analysis of the whole sentence fails. In the experiment, we detect gapping based on morphology and linear order of the words without using syntactic or semantic information. The test corpus, Finnish Wikipedia, is morphologically analyzed but not disambiguated. Even with an ambiguous morphological analysis, the results show that 89,9% of the detected sentences are elliptical, making the rules accurate enough to be used in the creation of FinnTreeBank. Once we have a morphologically disambiguated corpus, we can write more accurate rules and expect better results.

The Impact of Automatic Morphological Analysis & Disambiguation on Dependency Parsing of Turkish

Gülşen Eryiğit

The studies on dependency parsing of Turkish so far gave their results on the Turkish Dependency Treebank. This treebank consists of sentences where gold standard part-of-speech tags are manually assigned to each word and the words forming multi word expressions are also manually determined and combined into single units. For the first time, we investigate the results of parsing Turkish sentences from scratch and observe the accuracy drop at the end of processing raw data. We test one state-of-the art morphological analyzer together with two different morphological disambiguators. We both show separately the accuracy drop due to the automatic morphological processing and to the lack of multi word unit extraction. With this purpose, we use and present a new version of the Turkish Treebank where we detached the multi word expressions (MWEs) into multiple tokens and manually annotated the missing part-of-speech tags of these new tokens.

Task-Driven Linguistic Analysis based on an Underspecified Features Representation

Stasinos Konstantopoulos, Valia Kordoni, Nicola Cancedda, Vangelis Karkaletsis, Dietrich Klakow and Jean-Michel Renders

In this paper we explore a task-driven approach to interfacing NLP components, where language processing is guided by the end-task that each application requires. The core idea is to generalize feature values into feature value distributions, representing under-specified feature values, and to fit linguistic pipelines with a back-channel of specification requests through which subsequent components can declare to preceding ones the importance of narrowing the value distribution of particular features that are critical for the current task.

"Combining Language Resources Into A Grammar-Driven Swedish Parser"

Malin Ahlberg and Ramona Enache

This paper describes work on a rule-based, open-source parser for Swedish. The central component is a wide-coverage grammar implemented in the GF formalism (Grammatical Framework), a dependently typed grammar formalism based on Martin-Löf type theory. GF has strong support for multilinguality and has so far been used successfully for controlled languages and recent experiments have showed that it is also possible to use the framework for parsing unrestricted language. In addition to GF, we use two other main resources: the Swedish treebank Talbanken and the electronic lexicon SALDO. By combining the grammar with a lexicon extracted from SALDO we obtain a parser accepting all sentences described by the given rules. We develop and test this on examples from Talbanken. The resulting parser gives a full syntactic analysis of the input sentences. It will be highly reusable, freely available, and as GF provides libraries for compiling grammars to a number of programming languages, chosen parts of the the grammar may be used in various NLP applications.

The Icelandic Parsed Historical Corpus (IcePaHC)

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson and Joel Wallenberg

We describe the background for and building of IcePaHC, a one million word parsed historical corpus of Icelandic which has just

been finished. This corpus which is completely free and open contains fragments of 60 texts ranging from the late 12th century to the present. We describe the text selection and text collecting process and discuss the quality of the texts and their conversion to modern Icelandic spelling. We explain why we choose to use a phrase structure Penn style annotation scheme and briefly describe the syntactic anno-tation process. We also describe a spin-off project which is only in its beginning stages: a parsed historical corpus of Faroese. Finally, we advocate the importance of an open source policy as regards language resources.

A treebank-based study on the influence of Italian word order on parsing performance

Anita Alicante, Cristina Bosco, Anna Corazza and Alberto Lavelli

The aim of this paper is to contribute to the debate on the issues raised by Morphologically Rich Languages, and more precisely to investigate, in a cross-paradigm perspective, the influence of the constituent order on the data-driven parsing of one of such languages(i.e. Italian). It shows therefore new evidence from experiments on Italian, a language characterized by a rich verbal inflection, which leads to a widespread diffusion of the pro–drop phenomenon and to a relatively free word order. The experiments are performed by using state-of-the-art data-driven parsers (i.e. MaltParser and Berkeley parser) and are based on an Italian treebank available in formats that vary according to two dimensions, i.e. the paradigm of representation (dependency vs. constituency) and the level of detail of linguistic information.

Effort of Genre Variation and Prediction of System Performance

Dong Wang and Fei Xia

Domain adaptation is an important task in order for NLP systems to work well in real applications. There has been extensive research on this topic. In this paper, we address two issues that are related to domain adaptation. The first question is how much genre variation will affect NLP systems' performance. We investigate the effect of genre variation on the performance of three NLP tools, namely, word segmenter, POS tagger, and parser. We choose the Chinese Penn Treebank (CTB) as our corpus. The second question is how one can estimate NLP systems' performance when gold standard on the test data does not exist. To answer the question, we extend the prediction model in (Ravi et al., 2008) to provide prediction for word segmentation and POS tagging as well. Our experiments show that the predicted scores are close to the real scores when tested on the CTB data.

P21 - Information Extraction (2)

Thursday, May 24, 11:45

Chairperson: **Paul Buitelaar** Poster Session

Statistical Section Segmentation in Free-Text Clinical Records

Michael Tepper, Daniel Capurro, Fei Xia, Lucy Vanderwende and Meliha Yetisgen-Yildiz

Automatically segmenting and classifying clinical free text into sections is an important first step to automatic information retrieval, information extraction and data mining tasks, as it helps to ground the significance of the text within. In this work we describe our approach to automatic section segmentation of clinical records such as hospital discharge summaries and radiology reports, along with section classification into pre-defined section categories. We apply machine learning to the problems of section segmentation and section classification, comparing a joint (one-step) and a pipeline (two-step) approach. We demonstrate that our systems perform well when tested on three data sets, two for hospital discharge summaries and one for radiology reports. We then show the usefulness of section information by incorporating it in the task of extracting comorbidities from discharge summaries.

A Corpus of Scientific Biomedical Texts Spanning over 168 Years Annotated for Uncertainty

Ramona Bongelli, Carla Canestrari, Ilaria Riccioni, Andrzej Zuczkowski, Cinzia Buldorini, Ricardo Pietrobon, Alberto Lavelli and Bernardo Magnini

Uncertainty language permeates biomedical research and is fundamental for the computer interpretation of unstructured text. And yet, a coherent, cognitive-based theory to interpret Uncertainty language and guide Natural Language Processing is, to our knowledge, non-existing. The aim of our project was therefore to detect and annotate Uncertainty markers – which play a significant role in building knowledge or beliefs in readers' minds – in a biomedical research corpus. Our corpus includes 80 manually annotated articles from the British Medical Journal randomly sampled from a 168-year period. Uncertainty markers have been classified according to a theoretical framework based on a combined linguistic and cognitive theory. The corpus was manually annotated according to such principles. We performed preliminary experiments to assess the manually annotated corpus and establish a baseline for the automatic detection of Uncertainty markers. The results of the experiments show that most of the Uncertainty markers can be recognized with good accuracy.

Págico: Evaluating Wikipedia-based information retrieval in Portuguese

Cristina Mota, Alberto Simões, Cláudia Freitas, Luís Costa and Diana Santos

How do people behave in their everyday information seeking tasks, which often involve Wikipedia? Are there systems which can help them, or do a similar job? In this paper we describe Págico, an evaluation contest with the main purpose of fostering research in these topics. We describe its motivation, the collection of documents created, the evaluation setup, the topics chosen and their choice, the participation, as well as the measures used for evaluation and the gathered resources. The task–between information retrieval and question answering–can be further described as answering questions related to Portuguese-speaking culture in the Portuguese Wikipedia, in a number of different themes and geographic and temporal angles. This initiative allowed us to create interesting datasets and perform some assessment of Wikipedia, while also improving a public-domain open-source system for further wikipedia-based evaluations. In the paper, we provide examples of questions, we report the results obtained by the participants, and provide some discussion on complex issues.

Applying Random Indexing to Structured Data to Find Contextually Similar Words

Danica Damljanovic, Udo Kruschwitz, M-Dyaa Albakour, Johann Petrak and Mihai Lupu

Language resources extracted from structured data (e.g. Linked Open Data) have already been used in various scenarios to improve conventional Natural Language Processing techniques. The meanings of words and the relations between them are made more explicit in RDF graphs, in comparison to human-readable text, and hence have a great potential to improve legacy applications. In this paper, we describe an approach that can be used to extend or clarify the semantic meaning of a word by constructing a list of contextually related terms. Our approach is based on exploiting the structure inherent in an RDF graph and then applying the methods from statistical semantics, and in particular, Random Indexing, in order to discover contextually related terms. We evaluate our approach in the domain of life science using the dataset generated with the help of domain experts from a large pharmaceutical company (AstraZeneca). They were involved in two phases: firstly, to generate a set of keywords of interest to them, and secondly to judge the set of generated contextually similar words for each keyword of interest. We compare our proposed approach, exploiting the semantic graph, with the same method applied on the human readable text extracted from the graph.

The CONCISUS Corpus of Event Summaries

Horacio Saggion and Sandra Szasz

Text summarization and information extraction systems require adaptation to new domains and languages. This adaptation usually depends on the availability of language resources such as corpora. In this paper we present a comparable corpus in Spanish and English for the study of cross-lingual information extraction and summarization: the CONCISUS Corpus. It is a rich human-annotated dataset composed of comparable event summaries in Spanish and English covering four different domains: aviation accidents, rail accidents, earthquakes, and terrorist attacks. In addition to the monolingual summaries in English and Spanish, we provide automatic translations and “comparable” full event reports of the events. The human annotations are concepts marked in the textual sources representing the key event information associated to the event type. The dataset has also been annotated using text processing pipelines. It is being made freely available to the research community for research purposes.

Building and Exploring Semantic Equivalences Resources

Gracinda Carvalho, David Martins de Matos and Vitor Rocio

Language resources that include semantic equivalences at word level are common, and its usefulness is well established in text processing applications, as in the case of search. Named entities also play an important role for text based applications, but are not usually covered by the previously mentioned resources. The present work describes the WES base, Wikipedia Entity Synonym base, a freely available resource based on the Wikipedia. The WES base was built for the Portuguese Language, with the same format of another freely available thesaurus for the same language, the TeP base, which allows integration of equivalences both at word level and entity level. The resource has been built in a language independent way, so that it can be extended to different languages. The WES base was used in a Question Answering system, enhancing significantly its performance.

The TARSQI Toolkit

Marc Verhagen and James Pustejovsky

We present and demonstrate the updated version of the TARSQI Toolkit, a suite of temporal processing modules that extract temporal information from natural language texts. It parses the document and identifies temporal expressions, recognizes events,

anchor events to temporal expressions and orders events relative to each other. The toolkit was previously demonstrated at COLING 2008, but has since seen substantial changes including: (1) incorporation of a new time expression tagger, (2) embracement of stand-off annotation, (3) application to the medical domain and (4) introduction of narrative containers.

From medical language processing to BioNLP domain

Gabriella Pardelli, Manuela Sassi, Sara Goggi and Stefania Biagioni

This paper presents the results of a terminological work on a reference corpus in the domain of Biomedicine. In particular, the research tends to analyse the use of certain terms in Biomedicine in order to verify their change over the time with the aim of retrieving from the net the very essence of documentation. The terminological sample contains words used in BioNLP and biomedicine and identifies which terms are passing from scientific publications to the daily press and which are rather reserved to scientific production. The final scope of this work is to determine how scientific dissemination to an ever larger part of the society enables a public of common citizens to approach communication on biomedical research and development; and its main source is a reference corpus made up of three main repositories from which information related to BioNLP and Biomedicine is extracted. The paper is divided in three sections: 1) an introduction dedicated to data extracted from scientific documentation; 2) the second section devoted to methodology and data description; 3) the third part containing a statistical representation of terms extracted from the archive: indexes and concordances allow to reflect on the use of certain terms in this field and give possible keys for having access to the extraction of knowledge in the digital era.

Evaluation of a Complex Information Extraction Application in Specific Domain

Romaric Besançon, Olivier Ferret and Ludovic Jean-Louis

Operational intelligence applications in specific domains are developed using numerous natural language processing technologies and tools. A challenge for this integration is to take into account the limitations of each of these technologies in the global evaluation of the application. We present in this article a complex intelligence application for the gathering of information from the Web about recent seismic events. We present the different components needed for the development of such system, including Information Extraction, Filtering and Clustering, and the technologies behind each component. We also propose an independent evaluation of each component and an insight of their influence in the overall performance of the system.

A methodology for the extraction of information about the usage of formulaic expressions in scientific texts

Hannah Kermes

In this paper, we present a methodology for the extraction of formulaic expressions, which goes beyond the mere extraction of candidate patterns. Using a pipeline we are able to extract information about the usage of formulaic expressions automatically from text corpora. According to Biber and Barbieri (2007) formulaic expressions are “important building blocks of discourse in spoken and written registers”. The automatic extraction procedure can help to investigate the usage and function of these recurrent patterns in different registers and domains. Formulaic expressions are commonplace not only in every- day language but also in scientific writing. Patterns such as ’in this paper’, ’the number of’, ’on the basis of’ are often used by scientists to convey research interests, the theoretical basis of their studies, results of experiments, sci- entific findings as well as conclusions and are used as dis- course organizers. For Hyland (2008) they help to “shape meanings in specific context and contribute to our sense of coherence in a text”. We are interested in: (i) which and what type of formulaic expressions are used in scientific texts? (ii) the distribution of formulaic expression across different scien- tific disciplines, (iii) where do formulaic expressions occur within a text?

Structural alignment of plain text books

André Santos, José João Almeida and Nuno Carvalho

Text alignment is one of the main processes for obtaining parallel corpora. When aligning two versions of a book, results are often affected by unpaired sections – sections which only exist in one of the versions of the book. We developed Text::Perfide::BookSync, a Perl module which performs books synchronization (structural alignment based on section delimitation), provided they have been previously annotated by Text::Perfide::BookCleaner. We discuss the need for such a tool and several implementation decisions. The main functions are described, and examples of input and output are presented. Text::Perfide::PartialAlign is an extension of the partialAlign.py tool bundled with hunalign which proposes an alternative methods for splitting bitexts.

Dependency parsing for interaction detection in pharmacogenomics

Gerold Schneider, Fabio Rinaldi and Simon Clematide

We give an overview of our approach to the extraction of interactions between pharmacogenomic entities like drugs, genes and diseases and suggest classes of interaction types driven by data from PharmGKB and partly following the top level ontology WordNet and biomedical types from BioNLP. Our text mining approach to the extraction of interactions is based on syntactic analysis. We use syntactic analyses to explore domain events and to suggest a set of interaction labels for the pharmacogenomics domain.

A data and analysis resource for an experiment in text mining a collection of micro-blogs on a political topic.

William Black, Rob Procter, Steven Gray and Sophia Ananiadou

The analysis of a corpus of micro-blogs on the topic of the 2011 UK referendum about the Alternative Vote has been undertaken as a joint activity by text miners and social scientists. To facilitate the collaboration, the corpus and its analysis is managed in a Web-accessible framework that allows users to upload their own textual data for analysis and to manage their own text annotation resources used for analysis. The framework also allows annotations to be searched, and the analysis to be re-run after amending the analysis resources. The corpus is also doubly human-annotated stating both whether each tweet is overall positive or negative in sentiment and whether it is for or against the proposition of the referendum.

Invited Talk

Thursday, May 24, 13:10

Chairperson: **Mehmed Özkan**

The Turkish Language and its Challenges for Language Processing

Kemal Oflazer

Turkish is the language of over 70 million people in and around Turkey and the Turkic language family has over 150M speakers. Yet work on Turkish NLP has had a relatively short history. This talk aims to present an overview of aspects of Turkish that makes it interesting for NLP in general, and the challenges one faces as the language evolves relatively rapidly. The talk will cover the current state of Turkish NLP and computational resources available for the community.

O25 - Multimodal Corpora (2)

Thursday, May 24, 14:55

Chairperson: **Nick Campbell** Oral Session

SUTAV: A Turkish Audio-Visual Database

Ibrahim Saygin Topkaya and Hakan Erdogan

This paper contains information about the “Sabanci University Turkish Audio-Visual (SUTAV)” database. The main aim of collecting SUTAV database was to obtain a large audio-visual collection of spoken words, numbers and sentences in Turkish language. The database was collected between 2006 and 2010 during“Novel approaches in audio-visual speech recognition”project which is funded by The Scientific and Technological Research Council of Turkey (TUBITAK). First part of the database contains a large corpus of Turkish language and contains standart quality videos. The second part is relatively small compared to the first one and contains recordings of spoken digits in high quality videos. Although the main aim to collect SUTAV database was to obtain a database for audio-visual speech recognition applications, it also contains useful data that can be used in other kinds of multimodal research like biometric security and person verification. The paper presents information about the data collection process and the the spoken content. It also contains a sample evaluation protocol and recognition results that are obtained with a small portion of the database.

Multimodal Behaviour and Feedback in Different Types of Interaction

Costanza Navarretta and Patrizia Paggio

In this article, we compare feedback-related multimodal behaviours in two different types of interactions: first encounters between two participants who do not know each other in advance, and naturally-occurring conversations between two and three participants recorded at their homes. All participants are Danish native speakers. The interactions are transcribed using the same methodology, and the multimodal behaviours are annotated according to the same annotation scheme. In the study we focus on the most frequently occurring feedback expressions in the interactions and on feedback-related head movements and facial expressions. The analysis of the corpora, while confirming general facts about feedback-related head movements and facial expressions previously reported in the literature, also shows that the physical setting, the number of participants, the topics discussed, and the degree of familiarity influence the use of gesture types and the frequency of feedback-related expressions and gestures.

A Parallel Corpus of Music and Lyrics Annotated with Emotions

Carlo Strapparava, Rada Mihalcea and Alberto Battocchi

In this paper, we introduce a novel parallel corpus of music and lyrics, annotated with emotions at line level. We first describe the corpus, consisting of 100 popular songs, each of them including a music component, provided in the MIDI format, as well as a lyrics component, made available as raw text. We then describe our work on enhancing this corpus with emotion annotations using crowdsourcing. We also present some initial experiments on emotion classification using the music and the lyrics representations of the songs, which lead to encouraging results, thus demonstrating the promise of using joint music-lyric models for song processing.

Building a Multimodal Laughter Database for Emotion Recognition

Merlin Teodosia Suarez, Jocelynn Cu and Madelene Sta. Maria

Laughter is a significant paralinguistic cue that is largely ignored in multimodal affect analysis. In this work, we investigate how a multimodal laughter corpus can be constructed and annotated both with discrete and dimensional labels of emotions for acted and spontaneous laughter. Professional actors enacted emotions to produce acted clips, while spontaneous laughter was collected from volunteers. Experts annotated acted laughter clips, while volunteers who possess an acceptable empathic quotient score annotated spontaneous laughter clips. The data was pre-processed to remove noise from the environment, and then manually segmented starting from the onset of the expression until its offset. Our findings indicate that laughter carries distinct emotions, and that emotion in laughter is best recognized using audio information rather than facial information. This may be explained by emotion regulation, i.e. laughter is used to suppress or regulate certain emotions. Furthermore, contextual information plays a crucial role in understanding the kind of laughter and emotion in the enactment.

A Speech and Gesture Spatial Corpus in Assisted Living

Dimitra Anastasiou

Ambient Assisted Living (AAL) is the name for a European technology and innovation funding programme. AAL research field is about intelligent assistant systems for a healthier and safer life in the preferred living environments through the use of Information and Communication Technologies (ICT). We focus specifically on speech and gesture interaction which can enhance the quality of lifestyle of people living in assistive environments, be they seniors or people with physical or cognitive disabilities. In this paper we describe our user study conducted in a lab at the University of Bremen in order to collect empirical speech and gesture data and later create and analyse a multimodal corpus. The user study is about a human user sitting in a wheelchair and performing certain inherently spatial tasks.

O26 - Child Language Corpus

Thursday, May 24, 14:55

Chairperson: **Massimo Poesio** Oral Session

The Twins Corpus of Museum Visitor Questions

Priti Aggarwal, Ron Artstein, Jillian Gerten, Athanasios Katsamanis, Shrikanth Narayanan, Angela Nazarian and David Traum

The Twins corpus is a collection of utterances spoken in interactions with two virtual characters who serve as guides at the Museum of Science in Boston. The corpus contains about 200,000 spoken utterances from museum visitors (primarily children) as well as from trained handlers who work at the museum. In addition to speech recordings, the corpus contains the outputs of speech recognition performed at the time of utterance as well as the system interpretation of the utterances. Parts of the corpus have been manually transcribed and annotated for question interpretation. The corpus has been used for improving performance of the museum characters and for a variety of research projects, such as phonetic-based Natural Language Understanding, creation of conversational characters from text resources, dialogue policy learning, and research on patterns of user interaction. It has the potential to be used for research on children's speech and on language used when talking to a virtual human.

Korean Children's Spoken English Corpus and an Analysis of its Pronunciation Variability

Hyejin Hong, Sunhee Kim and Minhwa Chung

This paper introduces a corpus of Korean-accented English speech produced by children (the Korean Children's Spoken English Corpus: the KC-SEC), which is constructed by Seoul National University. The KC-SEC was developed in support of research and development of CALL systems for Korean learners of English, especially for elementary school learners. It consists of read-speech produced by 96 Korean learners aged from 9 to 12. Overall corpus size is 11,937 sentences, which amount to about 16 hours of speech. Furthermore, a statistical analysis of pronunciation variability appearing in the corpus is performed in

order to investigate the characteristics of the Korean children's spoken English. The realized phonemes (hypothesis) are extracted through time-based phoneme alignment, and are compared to the targeted phonemes (reference). The results of the analysis show that: i) the pronunciation variations found frequently in Korean children's speech are devoicing and changing of articulation place or/and manner; and ii) they largely correspond to those of general Korean learners' speech presented in previous studies, despite some differences.

Corpus of Children Voices for Mid-level Markers and Affect Bursts Analysis

Marie Tahon, Agnes Delaborde and Laurence Devillers

This article presents a corpus featuring children playing games in interaction with the humanoid robot Nao: children have to express emotions in the course of a storytelling by the robot. This corpus was collected to design an affective interactive system driven by an interactional and emotional representation of the user. We evaluate here some mid-level markers used in our system: reaction time, speech duration and intensity level. We also question the presence of affect bursts, which are quite numerous in our corpus, probably because of the young age of the children and the absence of predefined lexical content.

A large scale annotated child language construction database

Aline Villavicencio, Beracah Yankama, Marco Idiart and Robert Berwick

Large scale annotated corpora of child language can be of great value in assessing theoretical proposals regarding language acquisition models. For example, they can help determine whether the type and amount of data required by a proposed language acquisition model can actually be found in a naturalistic data sample. To this end, several recent efforts have augmented the CHILDES child language corpora with POS tagging and parsing information for languages such as English. With the increasing availability of robust NLP systems and electronic resources, these corpora can be further annotated with more detailed information about the properties of words, verb argument structure, and sentences. This paper describes such an initiative for combining information from various sources to extend the annotation of the English CHILDES corpora with linguistic, psycholinguistic and distributional information, along with an example illustrating an application of this approach to the extraction of verb alternation information. The end result, the English CHILDES Verb Construction Database, is an integrated resource containing information such as grammatical relations, verb semantic classes, and age of acquisition, enabling more targeted complex searches involving different levels of annotation that can facilitate a more detailed analysis of the linguistic input available to children.

Morphosyntactic Analysis of the CHILDES and TalkBank Corpora

Brian MacWhinney

This paper describes the construction and usage of the MOR and GRASP programs for part of speech tagging and syntactic dependency analysis of the corpora in the CHILDES and TalkBank databases. We have written MOR grammars for 11 languages and GRASP analyses for three. For English data, the MOR tagger reaches 98% accuracy on adult corpora and 97% accuracy on child language corpora. The paper discusses the construction of MOR lexicons with an emphasis on compounds and special conversational forms. The shape of rules for controlling allomorphy and morpheme concatenation are discussed. The analysis of bilingual corpora is illustrated in the context of the Cantonese-English bilingual corpora. Methods for preparing data for MOR analysis and for developing MOR grammars are discussed. We believe that recent computational work using this system is leading to significant advances in child language acquisition theory and theories of grammar identification more generally.

O27 - MultiWord Expressions

Thursday, May 24, 14:55

Chairperson: **Emanuele Pianta** Oral Session

Light Verb Constructions in the SzegedParalellFX English–Hungarian Parallel Corpus

Veronika Vincze

In this paper, we describe the first English-Hungarian parallel corpus annotated for light verb constructions, which contains 14,261 sentence alignment units. Annotation principles and statistical data on the corpus are also provided, and English and Hungarian data are contrasted. On the basis of corpus data, a database containing pairs of English-Hungarian light verb constructions has been created as well. The corpus and the database can contribute to the automatic detection of light verb constructions and it is also shown how they can enhance performance in several fields of NLP (e.g. parsing, information extraction/retrieval and machine translation).

Measuring the compositionality of NV expressions in Basque by means of distributional similarity techniques

Antton Gurrutxaga and Iñaki Alegria

We present several experiments aiming at measuring the semantic compositionality of NV expressions in Basque. Our approach

is based on the hypothesis that compositionality can be related to distributional similarity. The contexts of each NV expression are compared with the contexts of its corresponding components, by means of different techniques, as similarity measures usually used with the Vector Space Model (VSM), Latent Semantic Analysis (LSA) and some measures implemented in the Lemur Toolkit, as Indri index, tf-idf, Okapi index and Kullback-Leibler divergence. Using our previous work with cooccurrence techniques as a baseline, the results point to improvements using the Indri index or Kullback-Leibler divergence, and a slight further improvement when used in combination with cooccurrence measures such as t-score, via rank-aggregation. This work is part of a project for MWE extraction and characterization using different techniques aiming at measuring the properties related to idiomaticity, as institutionalization, non-compositionality and lexico-syntactic fixedness.

Analyzing and Aligning German compound nouns

Marion Weller and Ulrich Heid

In this paper, we present and evaluate an approach for the compositional alignment of compound nouns using comparable corpora from technical domains. The task of term alignment consists in relating a source language term to its translation in a list of target language terms with the help of a bilingual dictionary. Compound splitting allows to transform a compound into a sequence of components which can be translated separately and then related to multi-word target language terms. We present and evaluate a method for compound splitting, and compare two strategies for term alignment (bag-of-word vs. pattern-based). The simple word-based approach leads to a considerable amount of erroneous alignments, whereas the pattern-based approach reaches a decent precision. We also assess the reasons for alignment failures: in the comparable corpora used for our experiments, a substantial number of terms has no translation in the target language data; furthermore, the non-isomorphic structures of source and target language terms cause alignment failures in many cases.

Automatic Term Recognition Needs Multiple Evidence

Natalia Loukachevitch

In this paper we argue that the automatic term extraction procedure is an inherently multifactor process and the term extraction models needs to be based on multiple features including a specific type of a terminological resource under development. We proposed to use three types of features for extraction of two-word terms and showed that all these types of features are useful for term extraction. The set of features includes new features such as features extracted from an existing domain-specific thesaurus and features based on Internet search results. We studied the set of features for term extraction in two different domains and showed that the combination of several types of features considerably enhances the quality of the term extraction procedure. We found that for developing term extraction models in a specific domain, it is important to take into account some properties of the domain.

Constraint Based Description of Polish Multiword Expressions

Roman Kurc, Maciej Piasecki and Bartosz Broda

We present an approach to the description of Polish Multi-word Expressions (MWEs) which is based on expressions in the WCCL language of morpho-syntactic constraints instead of grammar rules or transducers. For each MWE its basic morphological form and the base forms of its constituents are specified but also each MWE is assigned to a class on the basis of its syntactic structure. For each class a WCCL constraint is defined which is parametrised by string variables referring to MWE constituent base forms or inflected forms. The constraint specifies a minimal set of conditions that must be fulfilled in order to recognise an occurrence of the given MWE in text with high accuracy. Our formalism is focused on the efficient description of large MWE lexicons for the needs of utilisation in text processing. The formalism allows for the relatively easy representation of flexible word order and discontinuous constructions. Moreover, there is no necessity for the full specification of the MWE grammatical structure. Only some aspects of the particular MWE structure can be selected in way facilitating the target accuracy of recognition. On the basis of a set of simple heuristics, WCCL-based representation of MWEs can be automatically generated from a list of MWE base forms. The proposed representation was applied on a practical scale for the description of a large set of Polish MWEs included in plWordNet.

O28 - Sign Language

Thursday, May 24, 14:55

Chairperson: **Javier Caminero** Oral Session

Recognition of Nonmanual Markers in American Sign Language (ASL) Using Non-Parametric Adaptive 2D-3D Face Tracking

Dimitris Metaxas, Bo Liu, Fei Yang, Peng Yang, Nicholas Michael and Carol Neidle

This paper addresses the problem of automatically recognizing linguistically significant nonmanual expressions in American Sign Language from video. We develop a fully automatic system

that is able to track facial expressions and head movements, and detect and recognize facial events continuously from video. The main contributions of the proposed framework are the following: (1) We have built a stochastic and adaptive ensemble of face trackers to address factors resulting in lost face track; (2) We combine 2D and 3D deformable face models to warp input frames, thus correcting for any variation in facial appearance resulting from changes in 3D head pose; (3) We use a combination of geometric features and texture features extracted from a canonical frontal representation. The proposed new framework makes it possible to detect grammatically significant nonmanual expressions from continuous signing and to differentiate successfully among linguistically significant expressions that involve subtle differences in appearance. We present results that are based on the use of a dataset containing 330 sentences from videos that were collected and linguistically annotated at Boston University.

Comparing computer vision analysis of signed language video with motion capture recordings

Matti Karppa, Tommi Jantunen, Ville Viitaniemi, Jorma Laaksonen, Birgitta Burger and Danny De Weerdt

We consider a non-intrusive computer-vision method for measuring the motion of a person performing natural signing in video recordings. The quality and usefulness of the method is compared to a traditional marker-based motion capture set-up. The accuracy of descriptors extracted from video footage is assessed qualitatively in the context of sign language analysis by examining if the shape of the curves produced by the different means resemble one another in sequences where the shape could be a source of valuable linguistic information. Then, quantitative comparison is performed first by correlating the computer-vision-based descriptors with the variables gathered with the motion capture equipment. Finally, multivariate linear and non-linar regression methods are applied for predicting the motion capture variables based on combinations of computer vision descriptors. The results show that even the simple computer vision method evaluated in this paper can produce promisingly good results for assisting researchers working on sign language analysis.

DEGELS1: A comparable corpus of French Sign Language and co-speech gestures

Annelies Braffort and Leïla Boutora

In this paper, we describe DEGELS1, a comparable corpus of French Sign Language and co-speech gestures that has been created to serve as a testbed corpus for the DEGELS workshops. These workshop series were initiated in France for researchers studying French Sign Language and co-speech gestures in French, with the aim of comparing methodologies for corpus annotation. An extract was used for the first event DEGELS2011 dedicated to the annotation of pointing, and the same extract will be used for DEGELS2012, dedicated to segmentation.

Semi-Automatic Sign Language Corpora Annotation using Lexical Representations of Signs

Matilde Gonzalez, Michael Filhol and Christophe Collet

Nowadays many researches focus on the automatic recognition of sign language. High recognition rates are achieved using lot of training data. This data is, generally, collected by manual annotating SL video corpus. However this is time consuming and the results depend on the annotators knowledge. In this work we intend to assist the annotation in terms of glosses which consist on writing down the sign meaning sign for sign thanks to automatic video processing techniques. In this case using learning data is not suitable since at the first step it will be needed to manually annotate the corpus. Also the context dependency of signs and the co-articulation effect in continuous SL make the collection of learning data very difficult. Here we present a novel approach which uses lexical representations of sign to overcome these problems and image processing techniques to match sign performances to sign representations. Signs are described using Zeebede (ZBD) which is a descriptor of signs that considers the high variability of signs. A ZBD database is used to stock signs and can be queried using several characteristics. From a video corpus sequence features are extracted using a robust body part tracking approach and a semi-automatic sign segmentation algorithm. Evaluation has shown the performances and limitation of the proposed approach.

A platform-independent user-friendly dictionary from Italian to LIS

Umar Shoaib, Nadeem Ahmad, Paolo Prinetto and Gabriele Tiotto

The Lack of written representation for Italian Sign Language (LIS) makes it difficult to do perform tasks like looking up a new word in a dictionary. Most of the paper dictionaries show LIS signs in drawings or pictures. It's not a simple proposition to understand the meaning of sign from paper dictionaries unless one already knows the meanings. This paper presents the LIS dictionary which provides the facility to translate Italian text into sign language. LIS signs are shown as video animations performed by a virtual character. The LIS dictionary provides the integration with MultiWordNet database. The integration with MultiWordNet allows a rich extension with the meanings and senses of the words existing in MultiWordNet. The dictionary allows users to acquire information about lemmas, synonyms and

synsets in the Sign Language (SL). The application is platform independent and can be used on any operating system. The results of input lemmas are displayed in groups of grammatical categories.

P22 - Part-of-Speech Tagging

Thursday, May 24, 14:55

Chairperson: **Reinhard Rapp** Poster Session

A Universal Part-of-Speech Tagset

Slav Petrov, Dipanjan Das and Ryan McDonald

To facilitate future research in unsupervised induction of syntactic structure and to standardize best-practices, we propose a tagset that consists of twelve universal part-of-speech categories. In addition to the tagset, we develop a mapping from 25 different treebank tagsets to this universal set. As a result, when combined with the original treebank data, this universal tagset and mapping produce a dataset consisting of common parts-of-speech for 22 different languages. We highlight the use of this resource via three experiments, that (1) compare tagging accuracies across languages, (2) present an unsupervised grammar induction approach that does not use gold standard part-of-speech tags, and (3) use the universal tags to transfer dependency parsers between languages, achieving state-of-the-art results.

Improving corpus annotation productivity: a method and experiment with interactive tagging

Atro Voutilainen

Corpus linguistic and language technological research needs empirical corpus data with nearly correct annotation and high volume to enable advances in language modelling and theorising. Recent work on improving corpus annotation accuracy presents semiautomatic methods to correct some of the analysis errors in available annotated corpora, while leaving the remaining errors undetected in the annotated corpus. We review recent advances in linguistics-based partial tagging and parsing, and regard the achieved analysis performance as sufficient for reconsidering a previously proposed method: combining nearly correct but partial automatic analysis with a minimal amount of human postediting (disambiguation) to achieve nearly correct corpus annotation accuracy at a competitive annotation speed. We report a pilot experiment with morphological (part-of-speech) annotation using a partial linguistic tagger of a kind previously reported with a very attractive precision-recall ratio, and observe that a desired level of annotation accuracy can be reached by using human disambiguation for less than 10% of the words in the corpus.

Lemmatising Serbian as Category Tagging with Bidirectional Sequence Classification

Andrea Gesmundo and Tanja Samardzic

We present a novel tool for morphological analysis of Serbian, which is a low-resource language with rich morphology. Our tool produces lemmatisation and morphological analysis reaching accuracy that is considerably higher compared to the existing alternative tools: 83.6% relative error reduction on lemmatisation and 8.1% relative error reduction on morphological analysis. The system is trained on a small manually annotated corpus with an approach based on Bidirectional Sequence Classification and Guided Learning techniques, which have recently been adapted with success to a broad set of NLP tagging tasks. In the system presented in this paper, this general approach to tagging is applied to the lemmatisation task for the first time thanks to our novel formulation of lemmatisation as a category tagging task. We show that learning lemmatisation rules from annotated corpus and integrating the context information in the process of morphological analysis provides a state-of-the-art performance despite the lack of resources. The proposed system can be used via a web GUI that deploys its best scoring configuration

Joint Segmentation and POS Tagging for Arabic Using a CRF-based Classifier

Souhir Gahbiche-Braham, Hélène Bonneau-Maynard, Thomas Lavergne and François Yvon

Arabic is a morphologically rich language, and Arabic texts abound of complex word forms built by concatenation of multiple subparts, corresponding for instance to prepositions, articles, roots prefixes, or suffixes. The development of Arabic Natural Language Processing applications, such as Machine Translation (MT) tools, thus requires some kind of morphological analysis. In this paper, we compare various strategies for performing such preprocessing, using generic machine learning techniques. The resulting tool is compared with two open domain alternatives in the context of a statistical MT task and is shown to be faster than its competitors, with no significant difference in MT quality.

Boosting statistical tagger accuracy with simple rule-based grammars

Mans Hulden and Jerid Francom

We report on several experiments on combining a rule-based tagger and a trigram tagger for Spanish. The results show that one can boost the accuracy of the best performing n-gram taggers by quickly developing a rough rule-based grammar to complement the statistically induced one and then combining the output of the two. The specific method of combination is crucial for achieving

good results. The method provides particularly large gains in accuracy when only a small amount of tagged data is available for training a HMM, as may be the case for lesser-resourced and minority languages.

NeoTag: a POS Tagger for Grammatical Neologism Detection

Maarten Janssen

POS Taggers typically fail to correctly tag grammatical neologisms: for known words, a tagger will only take known tags into account, and hence discard any possibility that the word is used in a novel or deviant grammatical category in the text at hand. Grammatical neologisms are relatively rare, and therefore do not pose a significant problem for the overall performance of a tagger. But for studies on neologisms and grammaticalization processes, this makes traditional taggers rather unfit. This article describes a modified POS tagger that explicitly considers new tags for known words, hence making it better fit for neologism research. This tagger, called NeoTag, has an overall accuracy that is comparable to other taggers, but scores much better for grammatical neologisms. To achieve this, the tagger applies a system of *lexical smoothing*, which adds new categories to known words based on known homographs. NeoTag also lemmatizes words as part of the tagging system, achieving a high accuracy on lemmatization for both known and unknown words, without the need for an external lexicon. The use of NeoTag is not restricted to grammatical neologism detection, and it can be used for other purposes as well.

Integrating NLP Tools in a Distributed Environment: A Case Study Chaining a Tagger with a Dependency Parser

Francesco Rubino, Francesca Frontini and Valeria Quochi

The present paper tackles the issue of PoS tag conversion within the framework of a distributed web service platform for the automatic creation of language resources. PoS tagging is now considered a“solved problem"; yet, because of the differences in the tagsets, interchange of the various PoS tagger available is still hampered. In this paper we describe the implementation of a pos-tagged-corpus converter, which is needed for chaining together in a workflow the Freeling PoS tagger for Italian and the DESR dependency parser, given that these two tools have been developed independently. The conversion problems experienced during the implementation, related to the properties of the different tagsets and of tagset conversion in general, are discussed together with the heuristics implemented in the attempt to solve them. Finally, the converter is evaluated by assessing the impact of conversion on the performance of the dependency parser. From this we learn that in most cases parsing errors are due to actual tagging errors, and not to conversion itself. Besides, information on accuracy loss is an important feature in a distributed environment of (NLP) services, where users need to decide which services best suit their needs.

P23 - Machine Translation (1)

Thursday, May 24, 14:55

Chairperson: **Philippe Langlais** Poster Session

Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies

Bruno Cartoni and Thomas Meyer

Translation studies rely more and more on corpus data to examine specificities of translated texts, that can be translated from different original languages and compared to original texts. In parallel, more and more multilingual corpora are becoming available for various natural language processing tasks. This paper questions the use of these multilingual corpora in translation studies and shows the methodological steps needed in order to obtain more reliably comparable sub-corpora that consist of original and directly translated text only. Various experiments are presented that show the advantage of directional sub-corpora.

Can Statistical Post-Editing with a Small Parallel Corpus Save a Weak MT Engine?

Marianna J. Martindale

Statistical post-editing has been shown in several studies to increase BLEU score for rule-based MT systems. However, previous studies have relied solely on BLEU and have not conducted further study to determine whether those gains indicated an increase in quality or in score alone. In this work we conduct a human evaluation of statistical post-edited output from a weak rule-based MT system, comparing the results with the output of the original rule-based system and a phrase-based statistical MT system trained on the same data. We show that for this weak rule-based system, despite significant BLEU score increases, human evaluators prefer the output of the original system. While this is not a generally conclusive condemnation of statistical post-editing, this result does cast doubt on the efficacy of statistical post-editing for weak MT systems and on the reliability of BLEU score for comparison between weak rule-based and hybrid systems built from them.

BLEU Evaluation of Machine-Translated English-Croatian Legislation

Sanja Seljan, Marija Brkić and Tomislav Vičić

This paper presents work on the evaluation of online available machine translation (MT) service, i.e. Google Translate, for English-Croatian language pair in the domain of legislation. The total set of 200 sentences, for which three reference translations are provided, is divided into short and long sentences. Human evaluation is performed by native speakers, using the criteria of adequacy and fluency. For measuring the reliability of agreement among raters, Fleiss' kappa metric is used. Human evaluation is enriched by error analysis, in order to examine the influence of error types on fluency and adequacy, and to use it in further research. Translation errors are divided into several categories: non-translated words, word omissions, unnecessarily translated words, morphological errors, lexical errors, syntactic errors and incorrect punctuation. The automatic evaluation metric BLEU is calculated with regard to a single and multiple reference translations. System level Pearson's correlation between BLEU scores based on a single and multiple reference translations is given, as well as correlation between short and long sentences BLEU scores, and correlation between the criteria of fluency and adequacy and each error category.

Chinese Characters Mapping Table of Japanese, Traditional Chinese and Simplified Chinese

Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi

Chinese characters are used both in Japanese and Chinese, which are called Kanji and Hanzi respectively. Chinese characters contain significant semantic information, a mapping table between Kanji and Hanzi can be very useful for many Japanese-Chinese bilingual applications, such as machine translation and cross-lingual information retrieval. Because Kanji characters are originated from ancient China, most Kanji have corresponding Chinese characters in Hanzi. However, the relation between Kanji and Hanzi is quite complicated. In this paper, we propose a method of making a Chinese characters mapping table of Japanese, Traditional Chinese and Simplified Chinese automatically by means of freely available resources. We define seven categories for Kanji based on the relation between Kanji and Hanzi, and classify mappings of Chinese characters into these categories. We use a resource from Wiktionary to show the completeness of the mapping table we made. Statistical comparison shows that our proposed method makes a more complete mapping table than the current version of Wiktionary.

Free/Open Source Shallow-Transfer Based Machine Translation for Spanish and Aragonese

Juan Pablo Martínez Cortés, Jim O'Regan and Francis Tyers

This article describes the development of a bidirectional shallow-transfer based machine translation system for Spanish and Aragonese, based on the Apertium platform, reusing the resources provided by other translators built for the platform. The system, and the morphological analyser built for it, are both the first resources of their kind for Aragonese. The morphological analyser has coverage of over 80%, and is being reused to create a spelling checker for Aragonese. The translator is bidirectional: the Word Error Rate for Spanish to Aragonese is 16.83%, while Aragonese to Spanish is 11.61%.

Automatic MT Error Analysis: Hjerson Helping Addicter

Jan Berka, Ondřej Bojar, Mark Fishel, Maja Popović and Daniel Zeman

We present a complex, open source tool for detailed machine translation error analysis providing the user with automatic error detection and classification, several monolingual alignment algorithms as well as with training and test corpus browsing. The tool is the result of a merge of automatic error detection and classification of Hjerson (Popović, 2011) and Addicter (Zeman et al., 2011) into the pipeline and web visualization of Addicter. It classifies errors into categories similar to those of Vilar et al. (2006), such as: morphological, reordering, missing words, extra words and lexical errors. The graphical user interface shows alignments in both training corpus and test data; the different classes of errors are colored. Also, the summary of errors can be displayed to provide an overall view of the MT system's weaknesses. The tool was developed in Linux, but it was tested on Windows too.

Re-ordering Source Sentences for SMT

Amit Sangodkar and Om Damani

We propose a pre-processing stage for Statistical Machine Translation (SMT) systems where the words of the source sentence are re-ordered as per the syntax of the target language prior to the alignment process, so that the alignment found by the statistical system is improved. We take a dependency parse of the source sentence and linearize it as per the syntax of the target language, before it is used in either the training or the decoding phase. During this linearization, the ordering decisions among dependency nodes having a common parent are done based on two aspects: parent-child positioning and relation priority. To

make the linearization process rule-driven, we assume that the relative word order of a dependency relation's relata does not depend either on the semantic properties of the relata or on the rest of the expression. We also assume that the relative word order of various relations sharing a relata does not depend on the rest of the expression. We experiment with a publicly available English-Hindi parallel corpus and show that our scheme improves the BLEU score.

An English-Portuguese parallel corpus of questions: translation guidelines and application in SMT

Ângela Costa, Tiago Luís, Joana Ribeiro, Ana Cristina Mendes and Luísa Coheur

The task of Statistical Machine Translation depends on large amounts of training corpora. Despite the availability of several parallel corpora, these are typically composed of declarative sentences, which may not be appropriate when the goal is to translate other types of sentences, e.g., interrogatives. There have been efforts to create corpora of questions, specially in the context of the evaluation of Question-Answering systems. One of those corpora is the UIUC dataset, composed of nearly 6,000 questions, widely used in the task of Question Classification. In this work, we make available the Portuguese version of the UIUC dataset, which we manually translated, as well as the translation guidelines. We show the impact of this corpus in the performance of a state-of-the-art SMT system when translating questions. Finally, we present a taxonomy of translation errors, according to which we analyze the output of the automatic translation before and after using the corpus as training data.

Word Alignment for English-Turkish Language Pair

Mehmet Talha Çakmak, Süleyman Acar and Gülşen Eryiğit

Word alignment is an important step for machine translation systems. Although the alignment performance between grammatically similar languages is reported to be very high in many studies, the case is not the same for language pairs from different language families. In this study, we are focusing on English-Turkish language pairs. Turkish is a highly agglutinative language with a very productive and rich morphology whereas English has a very poor morphology when compared to this language. As a result of this, one Turkish word is usually aligned with several English words. The traditional models which use word-level alignment approaches generally fail in such circumstances. In this study, we evaluate a Giza++ system by splitting the words into their morphological units (stem and suffixes) and compare the model with the traditional one. For the first time, we evaluate the performance of our aligner on gold standard parallel sentences rather than in a real machine translation system. Our approach reduced the alignment error rate by 40% relative. Finally, a new test corpus of 300 manually aligned sentences is released together with this study.

PEXACC: A Parallel Sentence Mining Algorithm from Comparable Corpora

Radu Ion

Extracting parallel data from comparable corpora in order to enrich existing statistical translation models is an avenue that attracted a lot of research in recent years. There are experiments that convincingly show how parallel data extracted from comparable corpora is able to improve statistical machine translation. Yet, the existing body of research on parallel sentence mining from comparable corpora does not take into account the degree of comparability of the corpus being processed or the computation time it takes to extract parallel sentences from a corpus of a given size. We will show that the performance of a parallel sentence extractor crucially depends on the degree of comparability such that it is more difficult to process a weakly comparable corpus than a strongly comparable corpus. In this paper we describe PEXACC, a distributed (running on multiple CPUs), trainable parallel sentence/phrase extractor from comparable corpora. PEXACC is freely available for download with the ACCURAT Toolkit, a collection of MT-related tools developed in the ACCURAT project.

A Richly Annotated, Multilingual Parallel Corpus for Hybrid Machine Translation

Eleftherios Avramidis, Marta R. Costa-Jussà, Christian Federmann, Josef van Genabith, Maite Melero and Pavel Pecina

In recent years, machine translation (MT) research has focused on investigating how hybrid machine translation as well as system combination approaches can be designed so that the resulting hybrid translations show an improvement over the individual "component" translations. As a first step towards achieving this objective we have developed a parallel corpus with source text and the corresponding translation output from a number of machine translation engines, annotated with metadata information, capturing aspects of the translation process performed by the different MT systems. This corpus aims to serve as a basic resource for further research on whether hybrid machine translation algorithms and system combination techniques can benefit from additional (linguistically motivated, decoding, and runtime) information provided by the different systems involved.

In this paper, we describe the annotated corpus we have created. We provide an overview on the component MT systems and the XLIFF-based annotation format we have developed. We also report on first experiments with the ML4HMT corpus data.

Automatic word alignment tools to scale production of manually aligned parallel texts

Stephen Grimes, Katherine Peterson and Xuansong Li

We have been creating large-scale manual word alignment corpora for Arabic-English and Chinese-English language pairs in genres such as newsire, broadcast news and conversation, and web blogs. We are now meeting the challenge of word aligning further varieties of web data for Chinese and Arabic “dialects". Human word alignment annotation can be costly and arduous. Alignment guidelines may be imprecise or underspecified in cases where parallel sentences are hard to compare – due to non-literal translations or differences between language structures. In order to speed annotation, we examine the effect that seeding manual alignments with automatic aligner output has on annotation speed and accuracy. We use automatic alignment methods that produce alignment results which are high precision and low recall to minimize annotator corrections. Results suggest that annotation time can be reduced by up to 20%, but we also found that reviewing and correcting automatic alignments requires more time than anticipated. We discuss throughout the paper crucial decisions on data structures for word alignment that likely have a significant impact on our results.

Design and compilation of a specialized Spanish-German parallel corpus

Carla Parra Escartín

This paper discusses the design and compilation of the TRIS corpus, a specialized parallel corpus of Spanish and German texts. It will be used for phraseological research aimed at improving statistical machine translation. The corpus is based on the European database of Technical Regulations Information System (TRIS), containing 995 original documents written in German and Spanish and their translations into Spanish and German respectively. This parallel corpus is under development and the first version with 97 aligned file pairs was released in the first META-NORD upload of metadata and resources in November 2011. The second version of the corpus, described in the current paper, contains 205 file pairs which have been completely aligned at sentence level, which account for approximately 1,563,000 words and 70,648 aligned sentence pairs.

A Distributed Resource Repository for Cloud-Based Machine Translation

Jörg Tiedemann, Dorte Haltrup Hansen, Lene Offersgaard, Sussi Olsen and Matthias Zumpe

In this paper, we present the architecture of a distributed resource repository developed for collecting training data for building customized statistical machine translation systems. The repository is designed for the cloud-based translation service integrated in the Let’sMT! platform which is about to be launched to the public. The system includes important features such as automatic import and alignment of textual documents in a variety of formats, a flexible database for meta-information using modern key-value stores and a grid-based backend for running off-line processes. The entire system is very modular and supports highly distributed setups to enable a maximum of flexibility and scalability. The system uses secure connections and includes an effective permission management to ensure data integrity. In this paper, we also take a closer look at the task of sentence alignment. The process of alignment is extremely important for the success of translation models trained on the platform. Alignment decisions significantly influence the quality of SMT engines.

P24 - Corpus Creation, Processing, Usage (1)

Thursday, May 24, 14:55

Chairperson: **Takenobu Tokunaga** Poster Session

Parallel Data, Tools and Interfaces in OPUS

Jörg Tiedemann

This paper presents the current status of OPUS, a growing language resource of parallel corpora and related tools. The focus in OPUS is to provide freely available data sets in various formats together with basic annotation to be useful for applications in computational linguistics, translation studies and cross-linguistic corpus studies. In this paper, we report about new data sets and their features, additional annotation tools and models provided from the website and essential interfaces and on-line services included in the project.

The Polish Sejm Corpus

Maciej Ogrodniczuk

This document presents the first edition of the Polish Sejm Corpus – a new specialized resource containing transcribed, automatically annotated utterances of the Members of Polish Sejm (lower chamber of the Polish Parliament). The corpus data encoding is inherited from the National Corpus of Polish and enhanced with session metadata and structure. The multi-layered stand-off annotation contains sentence- and token-level segmentation, disambiguated morphosyntactic information, syntactic words and groups resulting from shallow parsing and named entities. The paper also outlines several novel ideas for corpus preparation, e.g. the notion of a live corpus, constantly populated with new data or the concept of linking corpus data with external databases

to enrich content. Although initial statistical comparison of the resource with the balanced corpus of general Polish reveals substantial differences in language richness, the resource makes a valuable source of linguistic information as a large (300 M segments) collection of quasi-spoken data ready to be aligned with the audio/video recording of sessions, currently being made publicly available by Sejm.

From keystrokes to annotated process data: Enriching the output of Inputlog with linguistic information

Lieve Macken, Veronique Hoste, Marielle Leijten and Luuk Van Waes

Keystroke logging tools are a valuable aid to monitor written language production. These tools record all keystrokes, including backspaces and deletions together with timing information. In this paper we report on an extension to the keystroke logging program Inputlog in which we aggregate the logged process data from the keystroke (character) level to the word level. The logged process data are further enriched with different kinds of linguistic information: part-of-speech tags, lemmata, chunk boundaries, syllable boundaries and word frequency. A dedicated parser has been developed that distils from the logged process data word-level revisions, deleted fragments and final product data. The linguistically-annotated output will facilitate the linguistic analysis of the logged data and will provide a valuable basis for more linguistically-oriented writing process research. The set-up of the extension to Inputlog is largely language-independent. As proof-of-concept, the extension has been developed for English and Dutch. Inputlog is freely available for research purposes.

A Curated Database for Linguistic Research: The Test Case of Cimbrian Varieties

Maristella Agosti, Birgit Alber, Giorgio Maria Di Nunzio, Marco Dussin, Stefan Rabanus and Alessandra Tomaselli

In this paper we present the definition of a conceptual approach for the information space entailed by a multidisciplinary and collaborative project,“Cimbrian as a test case for synchronic and diachronic language variation”, which provides linguists with a test bed for formal hypotheses concerning human language. Aims of the project are to collect, digitize and tag linguistic data from the German variety of Cimbrian - spoken in three areas of northern Italy: Giazza (VR), Luserna (TN), and Roana (VI) - and to make available on-line a valuable and innovative linguistic resource for the in-depth study of Cimbrian. The task is addressed by a multidisciplinary team of linguists and computer scientists who, combining their competence, aim to make available new tools for linguistic analysis

Introducing the Reference Corpus of Contemporary Portuguese Online

Michel Généreux, Iris Hendrickx and Amália Mendes

We present our work in processing the Reference Corpus of Contemporary Portuguese and its publication online. After discussing how the corpus was built and our choice of meta-data, we turn to the processes and tools involved for the cleaning, preparation and annotation to make the corpus suitable for linguistic inquiries. The Web platform is described, and we show examples of linguistic resources that can be extracted from the platform for use in linguistic studies or in NLP.

A Basic Language Resource Kit for Persian

Mojgan Seraji, Beáta Megyesi and Joakim Nivre

Persian with its about 100,000,000 speakers in the world belongs to the group of languages with less developed linguistically annotated resources and tools. The few existing resources and tools are neither open source nor freely available. Thus, our goal is to develop open source resources such as corpora and treebanks, and tools for data-driven linguistic analysis of Persian. We do this by exploring the reusability of existing resources and adapting state-of-the-art methods for the linguistic annotation. We present fully functional tools for text normalization, sentence segmentation, tokenization, part-of-speech tagging, and parsing. As for resources, we describe the Uppsala PErsian Corpus (UPEC) which is a modified version of the Bijankhan corpus with additional sentence segmentation and consistent tokenization modified for more appropriate syntactic annotation. The corpus consists of 2,782,109 tokens and is annotated with parts of speech and morphological features. A treebank is derived from UPEC with an annotation scheme based on Stanford Typed Dependencies and is planned to consist of 10,000 sentences of which 215 have already been annotated. Keywords: BLARK for Persian, PoS tagged corpus, Persian treebank

Collecting and Analysing Chats and Tweets in SoNaR

Eric Sanders

In this paper a collection of chats and tweets from the Netherlands and Flanders is described. The chats and tweets are part of the freely available SoNaR corpus, a 500 million word text corpus of the Dutch language. Recruitment, metadata, anonymisation and IPR issues are discussed. To illustrate the difference of language use between the various text types and other parameters (like gender and age) simple text analysis in the form of unigram

frequency lists is carried out. Furthermore a website is presented with which users can retrieve their own frequency lists.

The goo300k corpus of historical Slovene

Tomaž Erjavec

The paper presents a gold-standard reference corpus of historical Slovene containing 1,000 sampled pages from over 80 texts, which were, for the most part, written between 1750-1900. Each page of the transcription has an associated facsimile and the words in the texts have been manually annotated with their modern-day equivalent, lemma and part-of-speech. The paper presents the structure of the text collection, the sampling procedure, annotation process and encoding of the corpus. The corpus is meant to facilitate HLT research and enable corpus based diachronic studies for historical Slovene. The corpus is encoded according to the Text Encoding Initiative Guidelines (TEI P5), is available via a concordancer and for download from http://nl.ijs.si/imp/ under the Creative Commons Attribution licence.

Kitten: a tool for normalizing HTML and extracting its textual content

Mathieu-Henri Falco, Véronique Moriceau and Anne Vilnat

The web is composed of a gigantic amount of documents that can be very useful for information extraction systems. Most of them are written in HTML and have to be rendered by an HTML engine in order to display the data they contain on a screen. HTML file thus mix both informational and rendering content. Our goal is to design a tool for informational content extraction. A linear extraction with only a basic filtering of rendering content would not be enough as objects such as lists and tables are linearly coded but need to be read in a non-linear way to be well interpreted. Besides these HTML pages are often incorrectly coded from an HTML point of view and use a segmentation of blocks based on blank space that cannot be transposed in a text filewithout confusing syntactic parsers. For this purpose, we propose the Kitten tool that first normalizes HTML file into unicode XHTML file, then extracts the informational content into a text filewith a special processing for sentences, lists and tables.

Collection of a corpus of Dutch SMS

Maaske Treurniet, Orphée De Clercq, Henk van den Heuvel and Nelleke Oostdijk

In this paper we present the first freely available corpus of Dutch text messages containing data originating from the Netherlands and Flanders. This corpus has been collected in the framework of the SoNaR project and constitutes a viable part of this 500-million-word corpus. About 53,000 text messages were collected on a large scale, based on voluntary donations. These messages will be distributed as such. In this paper we focus on the data collection processes involved and after studying the effect of media coverage we show that especially free publicity in newspapers and on social media networks results in more contributions. All SMS are provided with metadata information. Looking at the composition of the corpus, it becomes visible that a small number of people have contributed a large amount of data, in total 272 people have contributed to the corpus during three months. The number of women contributing to the corpus is larger than the number of men, but male contributors submitted larger amounts of data. This corpus will be of paramount importance for sociolinguistic research and normalisation studies.

RIDIRE-CPI: an Open Source Crawling and Processing Infrastructure for Supervised Web-Corpora Building

Alessandro Panunzi, Marco Fabbri, Massimo Moneglia, Lorenzo Gregori and Samuele Paladini

This paper introduces the RIDIRE-CPI, an open source tool for the building of web corpora with a specific design through a targeted crawling strategy. The tool has been developed within the RIDIRE Project, which aims at creating a 2 billion word balanced web corpus for Italian. RIDIRE-CPI architecture integrates existing open source tools as well as modules developed specifically within the RIDIRE project. It consists of various components: a robust crawler (Heritrix), a user friendly web interface, several conversion and cleaning tools, an anti-duplicate filter, a language guesser, and a PoS tagger. The RIDIRE-CPI user-friendly interface is specifically intended for allowing collaborative work performance by users with low skills in web technology and text processing. Moreover, RIDIRE-CPI integrates a validation interface dedicated to the evaluation of the targeted crawling. Through the content selection, metadata assignment, and validation procedures, the RIDIRE-CPI allows the gathering of linguistic data with a supervised strategy that leads to a higher level of control of the corpus contents. The modular architecture of the infrastructure and its open-source distribution will assure the reusability of the tool for other corpus building initiatives.

The Minho Quotation Resource

Brett Drury and José João Almeida

Direct quotations from business leaders can provide a rich sample of language which is in common use in the world of commerce. This language used by business leaders often uses: metaphors, euphemisms, slang, obscenities and invented words. In addition

the business lexicon is dynamic because new words or terms will gain popularity with businessmen whilst obsolete words will exit their common vocabulary. In addition to being a rich source of language direct quotations from business leaders can have ”real world” consequences. For example, Gerald Ratner nearly bankrupted his company with an infamous candid comment at an Institute of Directors meeting in 1993. Currently, there is no ”direct quotations from business leaders” resource freely available to the research community. The ”Minho Quotation Resource” captures the business lexicon with in excess of 500,000 quotations from individuals from the business world. The quotations were captured from October 2009 and April 2011. The resource is available in a searchable Lucene index and will be available for download in May 2012

Evaluating Query Languages for a Corpus Processing System

Elena Frick, Carsten Schnober and Piotr Bański

This paper documents a pilot study conducted as part of the development of a new corpus processing system at the Institut für Deutsche Sprache in Mannheim and in the context of the ISO TC37 SC4/WG6 activity on the suggested work item proposal “Corpus Query Lingua Franca”. We describe the first phase of our research: the initial formulation of functionality criteria for query language evaluation and the results of the application of these criteria to three representatives of corpus query languages, namely COSMAS II, Poliqarp, and ANNIS QL. In contrast to previous works on query language evaluation that compare a range of existing query languages against a small number of queries, our approach analyses only three query languages against criteria derived from a suite of 300 use cases that cover diverse aspects of linguistic research.

P25 - Evaluation Methodologies

Thursday, May 24, 14:55

Chairperson: **Mathieu Lafourcade** Poster Session

QurSim: A corpus for evaluation of relatedness in short texts

Abdul-Baquee Sharaf and Eric Atwell

This paper presents a large corpus created from the original Quranic text, where semantically similar or related verses are linked together. This corpus will be a valuable evaluation resource for computational linguists investigating similarity and relatedness in short texts. Furthermore, this dataset can be used for evaluation of paraphrase analysis and machine translation tasks. Our dataset is characterised by: (1) superior quality of relatedness assignment; as we have incorporated relations marked by well-known domain experts, this dataset could thus be considered a gold standard corpus for various evaluation tasks, (2) the size of our dataset; over 7,600 pairs of related verses are collected from scholarly sources with several levels of degree of relatedness. This dataset could be extended to over 13,500 pairs of related verses observing the commutative property of strongly related pairs. This dataset was incorporated into online query pages where users can visualize for a given verse a network of all directly and indirectly related verses. Empirical experiments showed that only 33% of related pairs shared root words, emphasising the need to go beyond common lexical matching methods, and incorporate -in addition- semantic, domain knowledge, and other corpus-based approaches.

EVALIEX – A Proposal for an Extended Evaluation Methodology for Information Extraction Systems

Christina Feilmayr, Birgit Pröll and Elisabeth Linsmayr

Assessing the correctness of extracted data requires performance evaluation, which is accomplished by calculating quality metrics. The evaluation process must cope with the challenges posed by information extraction and natural language processing. In the previous work most of the existing methodologies have been shown that they support only traditional scoring metrics. Our research work addresses requirements, which arose during the development of three productive rule-based information extraction systems. The main contribution is twofold: First, we developed a proposal for an evaluation methodology that provides the flexibility and effectiveness needed for comprehensive performance measurement. The proposal extends state-of-the-art scoring metrics by measuring string and semantic similarities and by parameterization of metric scoring, and thus simulating with human judgment. Second, we implemented an IE evaluation tool named EVALIEX, which integrates these measurement concepts and provides an efficient user interface that supports evaluation control and the visualization of IE results. To guarantee domain independence, the tool additionally provides a Generic Mapper for XML Instances (GeMap) that maps domain-dependent XML files containing IE results to generic ones. Compared to other tools, it provides more flexible testing and better visualization of extraction results for the comparison of different (versions of) information extraction systems.

A Rough Set Formalization of Quantitative Evaluation with Ambiguity

Patrick Paroubek and Xavier Tannier

In this paper, we present the founding elements of a formal model of the evaluation paradigm in natural language processing. We

propose an abstract model of objective quantitative evaluation based on rough sets, as well as the notion of potential performance space for describing the performance variations corresponding to the ambiguity present in hypothesis data produced by a computer program, when comparing it to the reference data created by humans. A formal model of the evaluation paradigm will be useful for comparing evaluations protocols, investigating evaluation constraint relaxation and getting a better understanding of the evaluation paradigm, provided it is general enough to be able to represent any natural language processing task.

The Influence of Corpus Quality on Statistical Measurements on Language Resources

Thomas Eckart, Uwe Quasthoff and Dirk Goldhahn

The quality of statistical measurements on corpora is strongly related to a strict definition of the measuring process and to corpus quality. In the case of multiple result inspections, an exact measurement of previously specified parameters ensures compatibility of the different measurements performed by different researchers on possibly different objects. Hence, the comparison of different values requires an exact description of the measuring process. To illustrate this correlation the influence of different definitions for the concepts“word”and“sentence”is shown for several properties of large text corpora. It is also shown that corpus pre-processing strongly influences corpus size and quality as well. As an example near duplicate sentences are identified as source of many statistical irregularities. The problem of strongly varying results especially holds for Web corpora with a large set of pre-processing steps. Here, a well-defined and language independent pre-processing is indispensable for language comparison based on measured values. Conversely, irregularities found in such measurements are often a result of poor pre-processing and therefore such measurements can help to improve corpus quality.

Identifying Nuggets of Information in GALE Distillation Evaluation

Olga Babko-Malaya, Greg Milette, Michael Schneider and Sarah Scogin

This paper describes an approach to automatic nuggetization and implemented system employed in GALE Distillation evaluation to measure the information content of text returned in response to an open-ended question. The system identifies nuggets, or atomic units of information, categorizes them according to their semantic type, and selects different types of nuggets depending on the type of the question. We further show how this approach addresses the main challenges for using automatic nuggetization for QA evaluation: the variability of relevant nuggets and their dependence on the question. Specifically, we propose a template-based approach to nuggetization, where different semantic categories of nuggets are extracted dependent on the template of a question. During evaluation, human annotators judge each snippet returned in response to a query as relevant or irrelevant, whereas automatic template-based nuggetization is further used to identify the semantic units of information that people would have selected as ‘relevant’ or ‘irrelevant’ nuggets for a given query. Finally, the paper presents the performance results of the nuggetization system which compare the number of automatically generated nuggets and human nuggets and show that our automatic nuggetization is consistent with human judgments.

NTUSocialRec: An Evaluation Dataset Constructed from Microblogs for Recommendation Applications in Social Networks

Chieh-Jen Wang, Shuk-Man Cheng, Lung-Hao Lee, Hsin-Hsi Chen, Wen-shen Liu, Pei-Wen Huang and Shih-Peng Lin

This paper proposes a method to construct an evaluation dataset from microblogs for the development of recommendation systems. We extract the relationships among three main entities in a recommendation event, i.e., who recommends what to whom. User-to-user friend relationships and user-to-resource interesting relationships in social media and resource-to-metadata descriptions in an external ontology are employed. In the experiments, the resources are restricted to visual entertainment media, movies in particular. A sequence of ground truths varying with time is generated. That reflects the dynamic of real world.

O29 - Language Generation and Paraphrasing

Thursday, May 24, 16:55

Chairperson: **Robert Dale** Oral Session

A contrastive review of paraphrase acquisition techniques

Houda Bouamor, Aurélien Max, Gabriel Illouz and Anne Vilnat

This paper addresses the issue of what approach should be used for building a corpus of sententential paraphrases depending on one’s requirements. Six strategies are studied: (1) multiple translations into a single language from another language; (2) multiple translations into a single language from different other languages; (3) multiple descriptions of short videos; (4) multiple subtitles for the same language; (5) headlines for similar news articles; and (6) sub-sentential paraphrasing in the context of a Web-based game.

We report results on French for 50 paraphrase pairs collected for all these strategies, where corpora were manually aligned at the finest possible level to define oracle performance in terms of accessible sub-sentential paraphrases. The differences observed will be used as criteria for motivating the choice of a given approach before attempting to build a new paraphrase corpus.

Chinese Whispers: Cooperative Paraphrase Acquisition

Matteo Negri, Yashar Mehdad, Alessandro Marchetti, Danilo Giampiccolo and Luisa Bentivogli

We present a framework for the acquisition of sentential paraphrases based on crowdsourcing. The proposed method maximizes the lexical divergence between an original sentence s and its valid paraphrases by running a sequence of paraphrasing jobs carried out by a crowd of non-expert workers. Instead of collecting direct paraphrases of s, at each step of the sequence workers manipulate semantically equivalent reformulations produced in the previous round. We applied this method to paraphrase English sentences extracted from Wikipedia. Our results show that, keeping at each round n the most promising paraphrases (i.e. the more lexically dissimilar from those acquired at round n-1), the monotonic increase of divergence allows to collect good-quality paraphrases in a cost-effective manner.

Diversifiable Bootstrapping for Acquiring High-Coverage Paraphrase Resource

Hideki Shima and Teruko Mitamura

Recognizing similar or close meaning on different surface form is a common challenge in various Natural Language Processing and Information Access applications. However, we identified multiple limitations in existing resources that can be used for solving the vocabulary mismatch problem. To this end, we will propose the Diversifiable Bootstrapping algorithm that can learn paraphrase patterns with a high lexical coverage. The algorithm works in a lightly-supervised iterative fashion, where instance and pattern acquisition are interleaved, each using information provided by the other. By tweaking a parameter in the algorithm, resulting patterns can be diversifiable with a specific degree one can control.

SemScribe: Natural Language Generation for Medical Reports

Sebastian Varges, Heike Bieler, Manfred Stede, Lukas C. Faulstich, Kristin Irsig and Malik Atalla

Natural language generation in the medical domain is heavily influenced by domain knowledge and genre-specific text characteristics. We present SemScribe, an implemented natural language generation system that produces doctor's letters, in particular descriptions of cardiological findings. Texts in this domain are characterized by a high density of information and a relatively telegraphic style. Domain knowledge is encoded in a medical ontology of about 80,000 concepts. The ontology is used in particular for concept generalizations during referring expression generation. Architecturally, the system is a generation pipeline that uses a corpus-informed syntactic frame approach for realizing sentences appropriate to the domain. The system reads XML documents conforming to the HL7 Clinical Document Architecture (CDA) Standard and enhances them with generated text and references to the used data elements. We conducted a first clinical trial evaluation with medical staff and report on the findings.

O30 - Computer Aided Language Learning

Thursday, May 24, 16:55

Chairperson: **Justus Roux** Oral Session

Item Development and Scoring for Japanese Oral Proficiency Testing

Hitokazu Matsushita and Deryle Lonsdale

This study introduces and evaluates a computerized approach to measuring Japanese L2 oral proficiency. We present a testing and scoring method that uses a type of structured speech called elicited imitation (EI) to evaluate accuracy of speech productions. Several types of language resources and toolkits are required to develop, administer, and score responses to this test. First, we present a corpus-based test item creation method to produce EI items with targeted linguistic features in a principled and efficient manner. Second, we sketch how we are able to bootstrap a small learner speech corpus to generate a significantly large corpus of training data for language model construction. Lastly, we show how newly created test items effectively classify learners according to their L2 speaking capability and illustrate how our scoring method computes a metric for language proficiency that correlates well with more traditional human scoring methods.

Evaluating Appropriateness Of System Responses In A Spoken CALL Game

Manny Rayner, Pierrette Bouillon and Johanna Gerlach

We describe an experiment carried out using a French version of CALL-SLT, a web-enabled CALL game in which students at each turn are prompted to give a semi-free spoken response which the system then either accepts or rejects. The central question we investigate is whether the response is appropriate; we do this by extracting pairs of utterances where both members of the pair are responses by the same student to the same prompt, and where one response is accepted and one rejected. When the two spoken responses are presented in random order, native speakers show a reasonable degree of agreement in judging that the accepted utterance is better than the rejected one. We discuss the significance of the results and also present a small study supporting the claim that native speakers are nearly always recognised by the system, while non-native speakers are rejected a significant proportion of the time.

Spontaneous Speech Corpora for language learners of Spanish, Chinese and Japanese

Antonio Moreno-Sandoval, Leonardo Campillos Llanos, Yang Dong, Emi Takamori, José M. Guirao, Paula Gozalo, Chieko Kimura, Kengo Matsui and Marta Garrote-Salazar

This paper presents a method for designing, compiling and annotating corpora intended for language learners. In particular, we focus on spoken corpora for being used as complementary material in the classroom as well as in examinations. We describe the three corpora (Spanish, Chinese and Japanese) compiled by the Laboratorio de Lingüística Informática at the Autonomous University of Madrid (LLI-UAM). A web-based concordance tool has been used to search for examples in the corpus, and providing the text along with the corresponding audio. Teaching materials from the corpus, consisting the texts, the audio files and exercises on them, are currently on development.

The DISCO ASR-based CALL system: practicing L2 oral skills and beyond

Helmer Strik, Jozef Colpaert, Joost Van Doremalen and Catia Cucchiarini

In this paper we describe the research that was carried out and the resources that were developed within the DISCO (Development and Integration of Speech technology into COurseware for language learning) project. This project aimed at developing an ASR-based CALL system that automatically detects pronunciation and grammar errors in Dutch L2 speaking and generates appropriate, detailed feedback on the errors detected. We briefly introduce the DISCO system and present its design, architecture and speech recognition modules. We then describe a first evaluation of the complete DISCO system and present some results. The resources generated through DISCO are subsequently described together with possible ways of efficiently generating additional resources in the future.

O31 - Discourse (2)

Thursday, May 24, 16:55

Chairperson: **Aravind Joshi** Oral Session

A Tool for Extracting Conversational Implicatures

Marta Tatu and Dan Moldovan

Explicitly conveyed knowledge represents only a portion of the information communicated by a text snippet. Automated mechanisms for deriving explicit information exist; however, the implicit assumptions and default inferences that capture our intuitions about a normal interpretation of a communication remain hidden for automated systems, despite the communication participants' ease of grasping the complete meaning of the communication. In this paper, we describe a reasoning framework for the automatic identification of conversational implicatures conveyed by real-world English and Arabic conversations carried via twitter.com. Our system transforms given utterances into deep semantic logical forms. It produces a variety of axioms that identify lexical connections between concepts, define rules of combining semantic relations, capture common-sense world knowledge, and encode Grice's Conversational Maxims. By exploiting this rich body of knowledge and reasoning within the context of the conversation, our system produces entailments and implicatures conveyed by analyzed utterances with an F-measure of 70.42% for English conversations.

Discourse-level Annotation over Europarl for Machine Translation: Connectives and Pronouns

Andrei Popescu-Belis, Thomas Meyer, Jeevanthi Liyanapathirana, Bruno Cartoni and Sandrine Zufferey

This paper describes methods and results for the annotation of two discourse-level phenomena, connectives and pronouns, over a multilingual parallel corpus. Excerpts from Europarl in English and French have been annotated with disambiguation information for connectives and pronouns, for about 3600 tokens. This data is then used in several ways: for cross-linguistic studies, for training automatic disambiguation software, and ultimately for training and testing discourse-aware statistical machine translation systems. The paper presents the annotation procedures and their

results in detail, and overviews the first systems trained on the annotated resources and their use for machine translation.

Annotating Story Timelines as Temporal Dependency Structures

Steven Bethard, Oleksandr Kolomiyets and Marie-Francine Moens

We present an approach to annotating timelines in stories where events are linked together by temporal relations into a temporal dependency tree. This approach avoids the disconnected timeline problems of prior work, and results in timelines that are more suitable for temporal reasoning. We show that annotating timelines as temporal dependency trees is possible with high levels of inter-annotator agreement - Krippendorff's Alpha of 0.822 on selecting event pairs, and of 0.700 on selecting temporal relation labels - even with the moderately sized relation set of BEFORE, AFTER, INCLUDES, IS-INCLUDED, IDENTITY and OVERLAP. We also compare several annotation schemes for identifying story events, and show that higher inter-annotator agreement can be reached by focusing on only the events that are essential to forming the timeline, skipping words in negated contexts, modal contexts and quoted speech.

An empirical resource for discovering cognitive principles of discourse organisation: the ANNODIS corpus

Stergos Afantenos, Nicholas Asher, Farah Benamara, Myriam Bras, Cecile Fabre, Mai Ho-Dac, Anne Le Draoulec, Philippe Muller, Marie-Paul Pery-Woodley, Laurent Prevot, Josette Rebeyrolles, Ludovic Tanguy, Marianne Vergez-Couret and Laure Vieu

This paper describes the ANNODIS resource, a discourse-level annotated corpus for French. The corpus combines two perspectives on discourse: a bottom-up approach and a top-down approach. The bottom-up view incrementally builds a structure from elementary discourse units, while the top-down view focuses on the selective annotation of multi-level discourse structures. The corpus is composed of texts that are diversified with respect to genre, length and type of discursive organisation. The methodology followed here involves an iterative design of annotation guidelines in order to reach satisfactory inter-annotator agreement levels. This allows us to raise a few issues relevant for the comparison of such complex objects as discourse structures. The corpus also serves as a source of empirical evidence for discourse theories. We present here two first analyses taking advantage of this new annotated corpus –one that tested hypotheses on constraints governing discourse structure, and another that studied the variations in composition and signalling of multi-level discourse structures.

O32 - Syntax and Parsing

Thursday, May 24, 16:55

Chairperson: **Adam Przepiórkowski** Oral Session

HamleDT: To Parse or Not to Parse?

Daniel Zeman, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský and Jan Hajič

We propose HamleDT – HArmonized Multi-LanguagE Dependency Treebank. HamleDT is a compilation of existing dependency treebanks (or dependency conversions of other treebanks), transformed so that they all conform to the same annotation style. While the license terms prevent us from directly redistributing the corpora, most of them are easily acquirable for research purposes. What we provide instead is the software that normalizes tree structures in the data obtained by the user from their original providers.

Evaluating and improving syntactic lexica by plugging them within a parser

Elsa Tolone, Benoît Sagot and Éric Villemonte de la Clergerie

We present some evaluation results for four French syntactic lexica, obtained through their conversion to the Alexina format used by the Lefff lexicon, and their integration within the large-coverage TAG-based FRMG parser. The evaluations are run on two test corpora, annotated with two distinct annotation formats, namely EASy/Passage chunks and relations and CoNLL dependencies. The information provided by the evaluation results provide valuable feedback about the four lexica. Moreover, when coupled with error mining techniques, they allow us to identify how these lexica might be improved.

Efficient Dependency Graph Matching with the IMS Open Corpus Workbench

Thomas Proisl and Peter Uhrig

State-of-the-art dependency representations such as the Stanford Typed Dependencies may represent the grammatical relations in a sentence as directed, possibly cyclic graphs. Querying a syntactically annotated corpus for grammatical structures that are represented as graphs requires graph matching, which is a non-trivial task. In this paper, we present an algorithm for graph matching that is tailored to the properties of large, syntactically annotated corpora. The implementation of the algorithm is built

on top of the popular IMS Open Corpus Workbench, allowing corpus linguists to re-use existing infrastructure. An evaluation of the resulting software, CWB-treebank, shows that its performance in real world applications, such as a web query interface, compares favourably to implementations that rely on a relational database or a dedicated graph database while at the same time offering a greater expressive power for queries. An intuitive graphical interface for building the query graphs is available via the Treebank.info project.

MaltOptimizer: A System for MaltParser Optimization

Miguel Ballesteros and Joakim Nivre

Freely available statistical parsers often require careful optimization to produce state-of-the-art results, which can be a non-trivial task especially for application developers who are not interested in parsing research for its own sake. We present MaltOptimizer, a freely available tool developed to facilitate parser optimization using the open-source system MaltParser, a data-driven parser-generator that can be used to train dependency parsers given treebank data. MaltParser offers a wide range of parameters for optimization, including nine different parsing algorithms, two different machine learning libraries (each with a number of different learners), and an expressive specification language that can be used to define arbitrarily rich feature models. MaltOptimizer is an interactive system that first performs an analysis of the training set in order to select a suitable starting point for optimization and then guides the user through the optimization of parsing algorithm, feature model, and learning algorithm. Empirical evaluation on data from the CoNLL 2006 and 2007 shared tasks on dependency parsing shows that MaltOptimizer consistently improves over the baseline of default settings and sometimes even surpasses the result of manual optimization.

P26 - Multilinguality

Thursday, May 24, 16:55

Chairperson: **Gil Francopoulo** Poster Session

Representing the Translation Relation in a Bilingual Wordnet

Jyrki Niemi and Krister Lindén

This paper describes representing translations in the Finnish wordnet, FinnWordNet (FiWN), and constructing the FiWN database. FiWN was created by translating all the word senses of the Princeton WordNet (PWN) into Finnish and by joining the translations with the semantic and lexical relations of PWN extracted into a relational (database) format. The approach naturally resulted in a translation relation between PWN and FiWN. Unlike many other multilingual wordnets, the translation relation in FiWN is not primarily on the synset level, but on the level of an individual word sense, which allows more precise translation correspondences. This can easily be projected into a synset-level translation relation, used for linking with other wordnets, for example, via Core WordNet. Synset-level translations are also used as a default in the absence of word-sense translations. The FiWN data in the relational database can be converted to other formats. In the PWN database format, translations are attached to source-language words, allowing the implementation of a Web search interface also working as a bilingual dictionary. Another representation encodes the translation relation as a finite-state transducer.

Building a multilingual parallel corpus for human users

Alexandr Rosen and Martin Vavřín

We present the architecture and the current state of InterCorp, a multilingual parallel corpus centered around Czech, intended primarily for human users and consisting of written texts with a focus on fiction. Following an outline of its recent development and a comparison with some other multilingual parallel corpora we give an overview of the data collection procedure that covers text selection criteria, data format, conversion, alignment, lemmatization and tagging. Finally, we show a sample query using the web-based search interface and discuss challenges and prospects of the project.

HunOr: A Hungarian–Russian Parallel Corpus

Martina Katalin Szabó, Veronika Vincze and István Nagy T.

In this paper, we present HunOr, the first multi-domain Hungarian–Russian parallel corpus. Some of the corpus texts have been manually aligned and split into sentences, besides, named entities also have been annotated while the other parts are automatically aligned at the sentence level and they are POS-tagged as well. The corpus contains texts from the domains literature, official language use and science, however, we would like to add texts from the news domain to the corpus. In the future, we are planning to carry out a syntactic annotation of the HunOr corpus, which will further enhance the usability of the corpus in various NLP fields such as transfer-based machine translation or cross lingual information retrieval.

Mining Hindi-English Transliteration Pairs from Online Hindi Lyrics

Kanika Gupta, Monojit Choudhury and Kalika Bali

This paper describes a method to mine Hindi-English transliteration pairs from online Hindi song lyrics. The technique is based on the observations that lyrics are transliterated word-by-word, maintaining the precise word order. The mining task is nevertheless challenging because the Hindi lyrics and its transliterations are usually available from different, often unrelated, websites. Therefore, it is a non-trivial task to match the Hindi lyrics to their transliterated counterparts. Moreover, there are various types of noise in lyrics data that needs to be appropriately handled before songs can be aligned at word level. The mined data of 30823 unique Hindi-English transliteration pairs with an accuracy of more than 92% is available publicly. Although the present work reports mining of Hindi-English word pairs, the same technique can be easily adapted for other languages for which song lyrics are available online in native and Roman scripts.

Dbnary: Wiktionary as a LMF based Multilingual RDF network

Gilles Sérasset

Contributive resources, such as wikipedia, have proved to be valuable in Natural Language Processing or Multilingual Information Retrieval applications.This article focusses on Wiktionary, the dictionary part of the collaborative resources sponsored by the Wikimedia foundation. In this article we present a word net that has been extracted from French, English and German wiktionaries. We present the structure of this word net and discuss the specific extraction problems induced by this kind of contributive resources and the method used to overcome them. Then we show how we represent the extracted data as a Lexical Markup Framework (LMF) compatible lexical network represented in Resource Description Framework (RDF) format.

FreeLing 3.0: Towards Wider Multilinguality

Lluís Padró and Evgeny Stanilovsky

FreeLing is an open-source multilingual language processing library providing a wide range of analyzers for several languages. It offers text processing and language annotation facilities to NLP application developers, lowering the cost of building those applications. FreeLing is customizable, extensible, and has a strong orientation to real-world applications in terms of speed and robustness. Developers can use the default linguistic resources (dictionaries, lexicons, grammars, etc.), extend/adapt them to specific domains, or –since the library is open source– develop new ones for specific languages or special application needs. This paper describes the general architecture of the library, presents the major changes and improvements included in FreeLing version 3.0, and summarizes some relevant industrial projects in which it has been used.

Bulgarian X-language Parallel Corpus

Svetla Koeva, Ivelina Stoyanova, Rositsa Dekova, Borislav Rizov and Angel Genov

The paper presents the methodology and the outcome of the compilation and the processing of the Bulgarian X-language Parallel Corpus (Bul-X-Cor) which was integrated as part of the Bulgarian National Corpus (BulNC). We focus on building representative parallel corpora which include a diversity of domains and genres, reflect the relations between Bulgarian and other languages and are consistent in terms of compilation methodology, text representation, metadata description and annotation conventions. The approaches implemented in the construction of Bul-X-Cor include using readily available text collections on the web, manual compilation (by means of Internet browsing) and preferably automatic compilation (by means of web crawling – general and focused). Certain levels of annotation applied to Bul-X-Cor are taken as obligatory (sentence segmentation and sentence alignment), while others depend on the availability of tools for a particular language (morpho-syntactic tagging, lemmatisation, syntactic parsing, named entity recognition, word sense disambiguation, etc.) or for a particular task (word and clause alignment). To achieve uniformity of the annotation we have either annotated raw data from scratch or transformed the already existing annotation to follow the conventions accepted for BulNC. Finally, actual uses of the corpora are presented and conclusions are drawn with respect to future work.

Automatically Generated Online Dictionaries

Enikő Héja and Dávid Takács

The aim of our software presentation is to demonstrate that corpus-driven bilingual dictionaries generated fully by automatic means are suitable for human use. Need for such dictionaries shows up specifically in the case of lesser used languages where due to the low demand it does not pay off for publishers to invest into the production of dictionaries. Previous experiments have proven that bilingual lexicons can be created by applying word alignment on parallel corpora. Such an approach, especially the corpus-driven nature of it, yields several advantages over more traditional approaches. Most importantly, automatically attained translation probabilities are able to guarantee that the most frequently used translations come first within an entry.

However, the proposed technique have to face some difficulties, as well. In particular, the scarce availability of parallel texts for medium density languages imposes limitations on the size of the resulting dictionary. Our objective is to design and implement a dictionary building workflow and a query system that is apt to exploit the additional benefits of the method and overcome the disadvantages of it.

Feedback in Nordic First-Encounters: a Comparative Study

Costanza Navarretta, Elisabeth Ahlsén, Jens Allwood, Kristiina Jokinen and Patrizia Paggio

The paper compares how feedback is expressed via speech and head movements in comparable corpora of first encounters in three Nordic languages: Danish, Finnish and Swedish. The three corpora have been collected following common guidelines, and they have been annotated according to the same scheme in the NOMCO project. The results of the comparison show that in this data the most frequent feedback-related head movement is Nod in all three languages. Two types of Nods were distinguished in all corpora: Down-nods and Up-nods; the participants from the three countries use Down- and Up-nods with different frequency. In particular, Danes use Down-nods more frequently than Finns and Swedes, while Swedes use Up-nods more frequently than Finns and Danes. Finally, Finns use more often single Nods than repeated Nods, differing from the Swedish and Danish participants. The differences in the frequency of both Down-nods and Up-Nods in the Danish, Finnish and Swedish interactions are interesting given that Nordic countries are not only geographically near, but are also considered to be very similar culturally. Finally, a comparison of feedback-related words in the Danish and Swedish corpora shows that Swedes and Danes use common feedback words corresponding to yes and no with similar frequency.

MultiUN v2: UN Documents with Multilingual Alignments

Yu Chen and Andreas Eisele

MultiUN is a multilingual parallel corpus extracted from the official documents of the United Nations. It is available in the six official languages of the UN and a small portion of it is also available in German. This paper presents a major update on the first public version of the corpus released in 2010. This version 2 consists of over 513,091 documents, including more than 9% of new documents retrieved from the United Nations official document system. We applied several modifications to the corpus preparation method. In this paper, we describe the methods we used for processing the UN documents and aligning the sentences. The most significant improvement compared to the previous release is the newly added multilingual sentence alignment information. The alignment information is encoded together with the text in XML instead of additional files. Our representation of the sentence alignment allows quick construction of aligned texts parallel in arbitrary number of languages, which is essential for building machine translation systems.

Customization of the Europarl Corpus for Translation Studies

Zahurul Islam and Alexander Mehler

Currently, the area of translation studies lacks corpora by which translation scholars can validate their theoretical claims, for example, regarding the scope of the characteristics of the translation relation. In this paper, we describe a customized resource in the area of translation studies that mainly addresses research on the properties of the translation relation. Our experimental results show that the Type-Token-Ratio (TTR) is not a universally valid indicator of the simplification of translation.

Accessing and standardizing Wiktionary lexical entries for the translation of labels in Cultural Heritage taxonomies

Thierry Declerck, Karlheinz Mörth and Piroska Lendvai

We describe the usefulness of Wiktionary, the freely available web-based lexical resource, in providing multilingual extensions to catalogues that serve content-based indexing of folktales and related narratives. We develop conversion tools between Wiktionary and TEI, using ISO standards (LMF, MAF), to make such resources available to both the Digital Humanities community and the Language Resources community. The converted data can be queried via a web interface, while the tools of the workflow are to be released with an open source license. We report on the actual state and functionality of our tools and analyse some shortcomings of Wiktionary, as well as potential domains of application.

A Mandarin-English Code-Switching Corpus

Ying Li, Yue Yu and Pascale Fung

Generally the existing monolingual corpora are not suitable for large vocabulary continuous speech recognition (LVCSR) of code-switching speech. The motivation of this paper is to study the rules and constraints code-switching follows and design a corpus for code-switching LVCSR task. This paper presents the development of a Mandarin-English code-switching corpus. This corpus consists of four parts: 1) conversational meeting speech and its data; 2) project meeting speech data; 3) student interviews speech; 4) text data of on-line news. The speech was transcribed

by an annotator and verified by Mandarin-English bilingual speakers manually. We propose an approach for automatically downloading from the web text data that contains code-switching. The corpus includes both intra-sentential code-switching (switch in the middle of a sentence) and inter-sentential code-switching (switch at the end of the sentence). The distribution of part-of-speech (POS) tags and code-switching reasons are reported.

A Fast, Memory Efficient, Scalable and Multilingual Dictionary Retriever

Paulo Fernandes, Lucelene Lopes, Carlos A. Prolo, Afonso Sales and Renata Vieira

This paper presents a novel approach to deal with dictionary retrieval. This new approach is based on a very efficient and scalable theoretical structure called Multi-Terminal Multi-valued Decision Diagrams (MTMDD). Such tool allows the definition of very large, even multilingual, dictionaries without significant increase in memory demands, and also with virtually no additional processing cost. Besides the general idea of the novel approach, this paper presents a description of the technologies involved, and their implementation in a software package called WAGGER. Finally, we also present some examples of usage and possible applications of this dictionary retriever.

Multilingual Central Repository version 3.0

Aitor Gonzalez-Agirre, Egoitz Laparra and German Rigau

This paper describes the upgrading process of the Multilingual Central Repository (MCR). The new MCR uses WordNet 3.0 as Interlingual-Index (ILI). Now, the current version of the MCR integrates in the same EuroWordNet framework wordnets from five different languages: English, Spanish, Catalan, Basque and Galician. In order to provide ontological coherence to all the integrated wordnets, the MCR has also been enriched with a disparate set of ontologies: Base Concepts, Top Ontology, WordNet Domains and Suggested Upper Merged Ontology. The whole content of the MCR is freely available.

P27 - Question Answering and Summarisation

Thursday, May 24, 16:55

Chairperson: **Horacio Saggion** Poster Session

A good space: Lexical predictors in word space evaluation

Christian Smith, Henrik Danielsson and Arne Jönsson

Vector space models benefit from using an outside corpus to train the model. It is, however, unclear what constitutes a good training corpus. We have investigated the effect on summary quality when using various language resources to train a vector space based extraction summarizer. This is done by evaluating the performance of the summarizer utilizing vector spaces built from corpora from different genres, partitioned from the Swedish SUC-corpus. The corpora are also characterized using a variety of lexical measures commonly used in readability studies. The performance of the summarizer is measured by comparing automatically produced summaries to human created gold standard summaries using the ROUGE F-score. Our results show that the genre of the training corpus does not have a significant effect on summary quality. However, evaluating the variance in the F-score between the genres based on lexical measures as independent variables in a linear regression model, shows that vector spaces created from texts with high syntactic complexity, high word variation, short sentences and few long words produce better summaries.

Creation and use of Language Resources in a Question-Answering eHealth System

Ulrich Andersen, Anna Braasch, Lina Henriksen, Csaba Huszka, Anders Johannsen, Lars Kayser, Bente Maegaard, Ole Norgaard, Stefan Schulz and Jürgen Wedekind

ESICT (Experience-oriented Sharing of health knowledge via Information and Communication Technology) is an ongoing research project funded by the Danish Council for Strategic Research. It aims at developing a health/disease related information system based on information technology, language technology, and formalized medical knowledge. The formalized medical knowledge consists partly of the terminology database SNOMED CT and partly of authorized medical texts on the domain. The system will allow users to ask questions in Danish and will provide natural language answers. Currently, the project is pursuing three basically different methods for question answering, and they are all described to some extent in this paper. A system prototype will handle questions related to diabetes and heart diseases. This paper concentrates on the methods employed for question answering and the language resources that are utilized. Some resources were existing, such as SNOMED CT, others, such as a corpus of sample questions, have had to be created or constructed.

Effects of Document Clustering in Modeling Wikipedia-style Term Descriptions

Atsushi Fujii, Yuya Fujii and Takenobu Tokunaga

Reflecting the rapid growth of science, technology, and culture, it has become common practice to consult tools on the World Wide Web for various terms. Existing search engines provide

an enormous volume of information, but retrieved information is not organized. Hand-compiled encyclopedias provide organized information, but the quantity of information is limited. In this paper, aiming to integrate the advantages of both tools, we propose a method to organize a search result based on multiple viewpoints as in Wikipedia. Because viewpoints required for explanation are different depending on the type of a term, such as animal and disease, we model articles in Wikipedia to extract a viewpoint structure for each term type. To identify a set of term types, we independently use manual annotation and automatic document clustering for Wikipedia articles. We also propose an effective feature for clustering of Wikipedia articles. We experimentally show that the document clustering reduces the cost for the manual annotation while maintaining the accuracy for modeling Wikipedia articles.

Evaluating Multi-focus Natural Language Queries over Data Services

Silvia Quarteroni, Vincenzo Guerrisi and Pietro La Torre

Natural language interfaces to data services will be a key technology to guarantee access to huge data repositories in an effortless way. This involves solving the complex problem of recognizing a relevant service or service composition given an ambiguous, potentially ungrammatical natural language question. As a first step toward this goal, we study methods for identifying the salient terms (or foci) in natural language questions, classifying the latter according to a taxonomy of services and extracting additional relevant information in order to route them to suitable data services. While current approaches deal with single-focus (and therefore single-domain) questions, we investigate multi-focus questions in the aim of supporting conjunctive queries over the data services they refer to. Since such complex queries have seldom been studied in the literature, we have collected an ad-hoc dataset, SeCo-600, containing 600 multi-domain queries annotated with a number of linguistic and pragmatic features. Our experiments with the dataset have allowed us to reach very high accuracy in different phases of query analysis, especially when adopting machine learning methods.

Summarizing a multimodal set of documents in a Smart Room

Maria Fuentes, Horacio Rodríguez and Jordi Turmo

This article reports an intrinsic automatic summarization evaluation in the scientific lecture domain. The lecture takes place in a Smart Room that has access to different types of documents produced from different media. An evaluation framework is presented to analyze the performance of systems producing summaries answering a user need. Several ROUGE metrics are used and a manual content responsiveness evaluation was carried out in order to analyze the performance of the evaluated approaches. Various multilingual summarization approaches are analyzed showing that the use of different types of documents outperforms the use of transcripts. In fact, not using any part of the spontaneous speech transcription in the summary improves the performance of automatic summaries. Moreover, the use of semantic information represented in the different textual documents coming from different media helps to improve summary quality.

P28 - Multimodal Corpus for Interaction

Thursday, May 24, 16:55

Chairperson: **Vangelis Karkaletsis** Poster Session

LAST MINUTE: a Multimodal Corpus of Speech-based User-Companion Interactions

Dietmar Rösner, Jörg Frommer, Rafael Friesen, Matthias Haase, Julia Lange and Mirko Otto

We report about design and characteristics of the LAST MINUTE corpus. The recordings in this data collection are taken from a WOZ experiment that allows to investigate how users interact with a companion system in a mundane situation with the need for planning, re-planning and strategy change. The resulting corpus is distinguished with respect to aspects of size (e.g. number of subjects, length of sessions, number of channels, total length of records) as well as quality (e.g. balancedness of cohort, well designed scenario, standard based transcripts, psychological questionnaires, accompanying in-depth interviews).

Annotating Football Matches: Influence of the Source Medium on Manual Annotation

Karën Fort and Vincent Claveau

In this paper, we present an annotation campaign of football (soccer) matches, from a heterogeneous text corpus of both match minutes and video commentary transcripts, in French. The data, annotations and evaluation process are detailed, and the quality of the annotated corpus is discussed. In particular, we propose a new technique to better estimate the annotator agreement when few elements of a text are to be annotated. Based on that, we show how the source medium influenced the process and the quality.

Creating HAVIC: Heterogeneous Audio Visual Internet Collection

Stephanie Strassel, Amanda Morris, Jonathan Fiscus, Christopher Caruso, Haejoong Lee, Paul Over, James Fiumara, Barbara Shaw, Brian Antonishek and Martial Michel

Linguistic Data Consortium and the National Institute of Standards and Technology are collaborating to create a large,

heterogeneous annotated multimodal corpus to support research in multimodal event detection and related technologies. The HAVIC (Heterogeneous Audio Visual Internet Collection) Corpus will ultimately consist of several thousands of hours of unconstrained user-generated multimedia content. HAVIC has been designed with an eye toward providing increased challenges for both acoustic and video processing technologies, focusing on multi-dimensional variation inherent in user-generated multimedia content. To date the HAVIC corpus has been used to support the NIST 2010 and 2011 TRECVID Multimedia Event Detection (MED) Evaluations. Portions of the corpus are expected to be released in LDC's catalog in the coming year, with the remaining segments being published over time after their use in the ongoing MED evaluations.

MULTIPHONIA: a MULTImodal database of PHONetics teaching methods in classroom InterActions.

Charlotte Alazard, Corine Astésano and Michel Billières

The Multiphonia Corpus consists of audio-video classroom recordings comparing two methods of phonetic correction (the 'traditional' articulatory method, and the Verbo-Tonal Method) This database was created not only to remedy the crucial lack of information and pedagogical resources on teaching pronunciation but also to test the benefit of VTM on Second Language pronunciation. The VTM method emphasizes the role of prosody cues as vectors of second language acquisition of the phonemic system. This method also provides various and unusual procedures including facilitating gestures in order to work on spotting and assimilating the target language prosodic system (rhythm, accentuation, intonation). In doing so, speech rhythm is apprehended in correlation with body/gestural rhythm. The student is thus encouraged to associate gestures activating the motor memory at play during the repetition of target words or phrases. In turn, pedagogical gestures have an impact on second language lexical items' recollection (Allen, 1995; Tellier, 2008). Ultimately, this large corpus (96 hours of class sessions' recordings) will be made available to the scientific community, with several layers of annotations available for the study of segmental, prosodic and gestural aspects of L2 speech.

P29 - Ontologies

Thursday, May 24, 16:55

Chairperson: **Paola Velardi** Poster Session

Mapping WordNet to the Kyoto ontology

Egoitz Laparra, German Rigau and Piek Vossen

This paper describes the connection of WordNet to a generic ontology based on DOLCE. We developed a complete set of heuristics for mapping all WordNet nouns, verbs and adjectives to the ontology. Moreover, the mapping also allows to represent predicates in a uniform and interoperable way, regardless of the way they are expressed in the text and in which language. Together with the ontology, the WordNet mappings provide a extremely rich and powerful basis for semantic processing of text in any domain. In particular, the mapping has been used in a knowledge-rich event-mining system developed for the Asian-European project KYOTO.

Constructing a Class-Based Lexical Dictionary using Interactive Topic Models

Kugatsu Sadamitsu, Kuniko Saito, Kenji Imamura and Yoshihiro Matsuo

This paper proposes a new method of constructing arbitrary class-based related word dictionaries on interactive topic models; we assume that each class is described by a topic. We propose a new semi-supervised method that uses the simplest topic model yielded by the standard EM algorithm; model calculation is very rapid. Furthermore our approach allows a dictionary to be modified interactively and the final dictionary has a hierarchical structure. This paper makes three contributions. First, it proposes a word-based semi-supervised topic model. Second, we apply the semi-supervised topic model to interactive learning; this approach is called the Interactive Topic Model. Third, we propose a score function; it extracts the related words that occupy the middle layer of the hierarchical structure. Experiments show that our method can appropriately retrieve the words belonging to an arbitrary class.

Adding Morpho-semantic Relations to the Romanian Wordnet

Verginica Barbu Mititelu

Keeping pace with other wordnets development, we present the challenges raised by the Romanian derivational system and our methodology for identifying derived words and their stems in the Romanian Wordnet. To attain this aim we rely only on the list of literals in the wordnet and on a list of Romanian affixes; the automatically obtained pairs require automatic and manual validation, based on a few heuristics. The correct members of the pairs are linked together and the relation is associated a semantic label whenever necessary. This label is proved to have cross-language validity. The work reported here contributes to the increase of the number of relations both between literals and between synsets, especially the cross-part-of-speech links. Words belonging to the same lexical family are identified easily. The benefits of thus improving a language resource such as wordnet become self-evident. The paper also contains an overview of the

current status of the Romanian wordnet and an envisaged plan for continuing the research.

An ontological approach to model and query multimodal concurrent linguistic annotations

Julien Seinturier, Elisabeth Murisasco, Emmanuel Bruno and Philippe Blache

This paper focuses on the representation and querying of knowledge-based multimodal data. This work stands in the OTIM project which aims at processing multimodal annotation of a large conversational French speech corpus. Within OTIM, we aim at providing linguists with a unique framework to encode and manipulate numerous linguistic domains (from prosody to gesture). Linguists commonly use Typed Feature Structures (TFS) to provide an uniform view of multimodal annotations but such a representation cannot be used within an applicative framework. Moreover TFS expressibility is limited to hierarchical and constituency relations and does not suit to any linguistic domain that needs for example to represent temporal relations. To overcome these limits, we propose an ontological approach based on Description logics (DL) for the description of linguistic knowledge and we provide an applicative framework based on OWL DL (Ontology Web Language) and the query language SPARQL.

The IMAGACT Cross-linguistic Ontology of Action. A new infrastructure for natural language disambiguation

Massimo Moneglia, Monica Monachini, Omar Calabrese, Alessandro Panunzi, Francesca Frontini, Gloria Gagliardi and Irene Russo

Action verbs, which are highly frequent in speech, cause disambiguation problems that are relevant to Language Technologies. This is a consequence of the peculiar way each natural language categorizes Action i.e. it is a consequence of semantic factors. Action verbs are frequently "general", since they extend productively to actions belonging to different ontological types. Moreover, each language categorizes action in its own way and therefore the cross-linguistic reference to everyday activities is puzzling. This paper briefly sketches the IMAGACT project, which aims at setting up a cross-linguistic Ontology of Action for grounding disambiguation tasks in this crucial area of the lexicon. The project derives information on the actual variation of action verbs in English and Italian from spontaneous speech corpora, where references to action are high in frequency. Crucially it makes use of the universal language of images to identify action types, avoiding the underdeterminacy of semantic definitions. Action concept entries are implemented as prototypic scenes; this will make it easier to extend the Ontology to other languages.

Towards a methodology for automatic identification of hypernyms in the definitions of large-scale dictionary

Inga Gheorghita and Jean-Marie Pierrel

The purpose of this paper is to identify automatically hypernyms for dictionary entries by exploring their definitions. In order to do this, we propose a weighting methodology that lets us assign to each lexeme a weight in a definition. This fact allows us to predict that lexemes with the highest weight are the closest hypernyms of the defined lexeme in the dictionary. The extracted semantic relation "is-a" is used for the automatic construction of a thesaurus for image indexing and retrieval. We conclude the paper by showing some experimental results to validate our method and by presenting our methodology of automatic thesaurus construction.

Collaborative semantic editing of linked data lexica

John McCrae, Elena Montiel-Ponsoda and Philipp Cimiano

The creation of language resources is a time-consuming process requiring the efforts of many people. The use of resources collaboratively created by non-linguistists can potentially ameliorate this situation. However, such resources often contain more errors compared to resources created by experts. For the particular case of lexica, we analyse the case of Wiktionary, a resource created along wiki principles and argue that through the use of a principled lexicon model, namely Lemon, the resulting data could be better understandable to machines. We then present a platform called Lemon Source that supports the creation of linked lexical data along the Lemon model. This tool builds on the concept of a semantic wiki to enable collaborative editing of the resources by many users concurrently. In this paper, we describe the model, the tool and present an evaluation of its usability based on a small group of users.

Ontoterminology: How to unify terminology and ontology into a single paradigm

Christophe Roche

Terminology is assigned to play a more and more important role in the Information Society. The need for a computational representation of terminology for IT applications raises new challenges for terminology. Ontology appears to be one of the most suitable solutions for such an issue. But an ontology is not a terminology as well as a terminology is not an

ontology. Terminology, especially for technical domains, relies on two different semiotic systems: the linguistic one, which is directly linked to the "Language for Special Purposes" and the conceptual system that describes the domain knowledge. These two systems must be both separated and linked. The new paradigm of ontoterminology, i.e. a terminology whose conceptual system is a formal ontology, emphasizes the difference between the linguistic and conceptual dimensions of terminology while unifying them. A double semantic triangle is introduced in order to link terms (signifiers) to concept names on a first hand and meanings (signified) to concepts on the other hand. Such an approach allows two kinds of definition to be introduced. The definition of terms written in natural language is considered as a linguistic explanation while the definition of concepts written in a formal language is viewed as a formal specification that allows operationalization of terminology.

Representation of linguistic and domain knowledge for second language learning in virtual worlds

Alexandre Denis, Ingrid Falk, Claire Gardent and Laura Perez-Beltrachini

There has been much debate, both theoretical and practical, on how to link ontologies and lexicons in natural language processing (NLP) applications. In this paper, we focus on an application in which lexicon and ontology are used to generate teaching material. We briefly describe the application (a serious game for language learning). We then zoom in on the representation and interlinking of the lexicon and of the ontology. We show how the use of existing standards and of good practice principles facilitates the design of our resources while satisfying the expressivity requirements set by natural language generation.

A Treebank-driven Creation of an OntoValence Verb lexicon for Bulgarian

Petya Osenova, Kiril Simov, Laska Laskova and Stanislava Kancheva

The paper presents a treebank-driven approach to the construction of a Bulgarian valence lexicon with ontological restrictions over the inner participants of the event. First, the underlying ideas behind the Bulgarian Ontology-based lexicon are outlined. Then, the extraction and manipulation of the valence frames is discussed with respect to the BulTreeBank annotation scheme and DOLCE ontology. Also, the most frequent types of syntactic frames are specified as well as the most frequent types of ontological restrictions over the verb arguments. The envisaged application of such a lexicon would be: in assigning ontological labels to syntactically parsed corpora, and expanding the lexicon and lexical information in the Bulgarian Resource Grammar.

Creation of a bottom-up corpus-based ontology for Italian Linguistics

Elisa Bianchi, Mirko Tavosanis and Emiliano Giovannetti

This paper describes the steps of construction of a shallow lexical ontology of Italian Linguistics, set to be used by a meta-search engine for query refinement. The ontology was constructed with the software Protégé 4.0.2 and is in OWL format; its construction has been carried out following the steps described in the well-known Ontology Learning From Text (OLFT) layer cake. The starting point was the automatic term extraction from a corpus of web documents concerning the domain of interest (304,000 words); as regards corpus construction, we describe the main criteria of the web documents selection and its critical points, concerning the definition of user profile and of degrees of specialisation. We describe then the process of term validation and construction of a glossary of terms of Italian Linguistics; afterwards, we outline the identification of synonymic chains and the main criteria of ontology design: top classes of ontology are Concept (containing taxonomy of concepts) and Terms (containing terms of the glossary as instances), while concepts are linked through part-whole and involved-role relation, both borrowed from Wordnet. Finally, we show some examples of the application of the ontology for query refinement.

Visualizing word senses in WordNet Atlas

Matteo Abrate and Clara Bacciu

This demo presents the second prototype of WordNet Atlas, a web application that gives users the ability to navigate and visualize the 146,312 word senses of the nouns contained within the Princeton WordNet. Two complementary, interlinked visualizations are provided: an hypertextual dictionary to represent detailed information about a word sense, such as lemma, definition and depictions, and a zoomable map representing the taxonomy of noun synsets in a circular layout. The application could help users unfamiliar with WordNet to get oriented in the large amount of data it contains.

O33 - Semantics from Corpora

Thursday, May 24, 18:20

Chairperson: **Maria Teresa Pazienza** Oral Session

Concept-based Selectional Preferences and Distributional Representations from Wikipedia Articles

Alex Judea, Vivi Nastase and Michael Strube

This paper describes the derivation of distributional semantic representations for open class words relative to a concept

inventory, and of concepts relative to open class words through grammatical relations extracted from Wikipedia articles. The concept inventory comes from WikiNet, a large-scale concept network derived from Wikipedia. The distinctive feature of these representations are their relation to a concept network, through which we can compute selectional preferences of open-class words relative to general concepts. The resource thus derived provides a meaning representation that complements the relational representation captured in the concept network. It covers English open-class words, but the concept base is language independent. The resource can be extended to other languages, with the use of language specific dependency parsers. Good results in metonymy resolution show the resource's potential use for NLP applications.

Associative and Semantic Features Extracted From Web-Harvested Corpora

Elias Iosif, Maria Giannoudaki, Eric Fosler-Lussier and Alexandros Potamianos

We address the problem of automatic classification of associative and semantic relations between words, and particularly those that hold between nouns. Lexical relations such as synonymy, hypernymy/hyponymy, constitute the fundamental types of semantic relations. Associative relations are harder to define, since they include a long list of diverse relations, e.g.,"Cause-Effect","Instrument-Agency". Motivated by findings from the literature of psycholinguistics and corpus linguistics, we propose features that take advantage of general linguistic properties. For evaluation we merged three datasets assembled and validated by cognitive scientists. A proposed priming coefficient that measures the degree of asymmetry in the order of appearance of the words in text achieves the best classification results, followed by context-based similarity metrics. The web-based features achieve classification accuracy that exceeds 85%.

Buildind a Resource of Patterns Using Semantic Types

Octavian Popescu

While a word in isolation has a high potential of expressing various senses, in certain phrases this potential is restricted up to the point that one and only one sense is possible. A phrase is called sense stable if the senses of all the words compounding it do not change their sense irrespective of the context which could be added to its left or to its right. By comparing sense stable phrases we can extract corpus patterns. These patterns have slots which are filled by semantic types that capture the relevant information for disambiguation. The relationship between slots is such that a chain like disambiguation process is possible. Annotating a corpus with these kinds of patterns is beneficial for NLP, because problems such as data sparseness, noise, learning complexity are alleviated. We evaluate the inter agreement of annotators on examples coming from BNC.

O34 - Authoring and Related Tools

Thursday, May 24, 18:20

Chairperson: **Bernardo Magnini** Oral Session

CLCM - A Linguistic Resource for Effective Simplification of Instructions in the Crisis Management Domain and its Evaluations

Irina Temnikova, Constantin Orasan and Ruslan Mitkov

Due to the increasing number of emergency situations which can have substantial consequences, both financially and fatally, the Crisis Management (CM) domain is developing at an exponential speed. The efficient management of emergency situations relies on clear communication between all of the participants in a crisis situation. For these reasons the Text Complexity (TC) of the CM domain needed to be investigated and showed that CM domain texts exhibit high TC levels. This article presents a new linguistic resource in the form of Controlled Language (CL) guidelines for manual text simplification in the CM domain which aims to address high TC in the CM domain and produce clear messages to be used in crisis situations. The effectiveness of the resource has been tested via evaluation from several different perspectives important for the domain. The overall results show that the CLCM simplification has a positive impact on TC, reading comprehension, manual translation and machine translation. Additionally, an investigation of the cognitive difficulty in applying manual simplification operations led to interesting discoveries. This article provides details of the evaluation methods, the conducted experiments, their results and indications about future work.

A Framework for Evaluating Text Correction

Robert Dale and George Narroway

Computer-based aids for writing assistance have been around since at least the early 1980s, focussing primarily on aspects such as spelling, grammar and style. The potential audience for such tools is very large indeed, and this is a clear case where we might expect to see language processing applications having a significant real-world impact. However, existing comparative evaluations of applications in this space are often no more than impressionistic and anecdotal reviews of commercial offerings as found in software magazines, making it hard to determine which approaches are superior. More rigorous evaluation in the

scholarly literature has been held back in particular by the absence of shared datasets of texts marked-up with errors, and the lack of an agreed evaluation framework. Significant collections of publicly available data are now appearing; this paper describes a complementary evaluation framework, which has been piloted in the Helping Our Own shared task. The approach, which uses stand-off annotations for representing edits to text, can be used in a wide variety of text-correction tasks, and easily accommodates different error tagsets.

Typing Race Games as a Method to Create Spelling Error Corpora

Paul Rodrigues and C. Anton Rytting

This paper presents a method to elicit spelling error corpora using an online typing race game. After being tested for their native language, English-native participants were instructed to retype stimuli as quickly and as accurately as they could. The participants were informed that the system was keeping a score based on accuracy and speed, and that a high score would result in a position on a public scoreboard. Words were presented on the screen one at a time from a queue, and the queue was advanced by pressing the ENTER key following the stimulus. Responses were recorded and compared to the original stimuli. Responses that differed from the stimuli were considered a typographical or spelling error, and added to an error corpus. Collecting a corpus using a game offers several unique benefits. 1) A game attracts engaged participants, quickly. 2) The web-based delivery reduces the cost and decreases the time and effort of collecting the corpus. 3) Participants have fun. Spelling error corpora have been difficult and expensive to obtain for many languages and this research was performed to fill this gap. In order to evaluate the methodology, we compare our game data against three existing spelling corpora for English.

O35 -Word Sense Annotation and Disambiguation

Thursday, May 24, 18:20

Chairperson: **Yoshihiko Hayashi** Oral Session

The MASC Word Sense Corpus

Rebecca J. Passonneau, Collin F. Baker, Christiane Fellbaum and Nancy Ide

The MASC project has produced a multi-genre corpus with multiple layers of linguistic annotation, together with a sentence corpus containing WordNet 3.1 sense tags for 1000 occurrences of each of 100 words produced by multiple annotators, accompanied by indepth inter-annotator agreement data. Here we give an overview of the contents of MASC and then focus on the word sense sentence corpus, describing the characteristics that differentiate it from other word sense corpora and detailing the inter-annotator agreement studies that have been performed on the annotations. Finally, we discuss the potential to grow the word sense sentence corpus through crowdsourcing and the plan to enhance the content and annotations of MASC through a community-based collaborative effort.

Addressing polysemy in bilingual lexicon extraction from comparable corpora

Darja Fišer, Nikola Ljubešić and Ozren Kubelka

This paper presents an approach to extract translation equivalents from comparable corpora for polysemous nouns. As opposed to the standard approaches that build a single context vector for all occurrences of a given headword, we first disambiguate the headword with third-party sense taggers and then build a separate context vector for each sense of the headword. Since state-of-the-art word sense disambiguation tools are still far from perfect, we also tried to improve the results by combining the sense assignments provided by two different sense taggers. Evaluation of the results shows that we outperform the baseline (0.473) in all the settings we experimented with, even when using only one sense tagger, and that the best-performing results are indeed obtained by taking into account the intersection of both sense taggers (0.720).

Empirical Comparisons of MASC Word Sense Annotations

Gerard de Melo, Collin F. Baker, Nancy Ide, Rebecca J. Passonneau and Christiane Fellbaum

We analyze how different conceptions of lexical semantics affect sense annotations and how multiple sense inventories can be compared empirically, based on annotated text. Our study focuses on the MASC project, where data has been annotated using WordNet sense identifiers on the one hand, and FrameNet lexical units on the other. This allows us to compare the sense inventories of these lexical resources empirically rather than just theoretically, based on their glosses, leading to new insights. In particular, we compute contingency matrices and develop a novel measure, the Expected Jaccard Index, that quantifies the agreement between annotations of the same data based on two different resources even when they have different sets of categories.

O36 - Time and Space

Thursday, May 24, 18:20

Chairperson: **Piek Vossen** Oral Session

TIMEN: An Open Temporal Expression Normalisation Resource

Hector Llorens, Leon Derczynski, Robert Gaizauskas and Estela Saquete

Temporal expressions are words or phrases that describe a point, duration or recurrence in time. Automatically annotating these expressions is a research goal of increasing interest. Recognising them can be achieved with minimally supervised machine learning, but interpreting them accurately (normalisation) is a complex task requiring human knowledge. In this paper, we present TIMEN, a community-driven tool for temporal expression normalisation. TIMEN is derived from current best approaches and is an independent tool, enabling easy integration in existing systems. We argue that temporal expression normalisation can only be effectively performed with a large knowledge base and set of rules. Our solution is a framework and system with which to capture this knowledge for different languages. Using both existing and newly-annotated data, we present results showing competitive performance and invite the IE community to contribute to a knowledge base in order to solve the temporal expression normalisation problem.

Annotating Spatial Containment Relations Between Events

Kirk Roberts, Travis Goodwin and Sanda M. Harabagiu

A significant amount of spatial information in textual documents is hidden within the relationship between events. While humans have an intuitive understanding of these relationships that allow us to recover an object's or event's location, currently no annotated data exists to allow automatic discovery of spatial containment relations between events. We present our process for building such a corpus of manually annotated spatial relations between events. Events form complex predicate-argument structures that model the participants in the event, their roles, as well as the temporal and spatial grounding. In addition, events are not presented in isolation in text; there are explicit and implicit interactions between events that often participate in event structures. In this paper, we focus on five spatial containment relations that may exist between events: (1) SAME, (2) CONTAINS, (3) OVERLAPS, (4) NEAR, and (5) DIFFERENT. Using the transitive closure across these spatial relations, the implicit location of many events and their participants can be discovered. We discuss our annotation schema for spatial containment relations, placing it within the pre-existing theories of spatial representation. We also discuss our annotation guidelines for maintaining annotation quality as well as our process for augmenting SpatialML with spatial containment relations between events. Additionally, we outline some baseline experiments to evaluate the feasibility of developing supervised systems based on this corpus. These results indicate that although the task is challenging, automated methods are capable of discovering spatial containment relations between events.

The Role of Model Testing in Standards Development: The Case of ISO-Space

James Pustejovsky and Jessica Moszkowicz

In this paper, we describe the methodology being used to develop certain aspects of ISO-Space, an annotation language for encoding spatial and spatiotemporal information as expressed in natural language text. After reviewing the requirements of a specification for capturing such knowledge from linguistic descriptions, we describe how ISO-Space has developed to meet the needs of the specification. ISO-Space is an emerging resource that is being developed in the context of an iterative effort to test the specification model with annotation, a methodology called MAMA (Model-Annotate-Model-Annotate) (Pustejovsky and Stubbs, 2012). We describe the genres of text that are being used in a pilot annotation study, in order to both refine and enrich the specification language by way of crowd sourcing simple annotation tasks with Amazon's Mechanical Turk Service.

P30 - Discourse

Thursday, May 24, 18:20

Chairperson: **David Traum** Poster Session

Investigating Engagement - intercultural and technological aspects of the collection, analysis, and use of the Estonian Multiparty Conversational video data

Kristiina Jokinen and Silvi Tenjes

In this paper we describe the goals of the Estonian corpus collection and analysis activities, and introduce the recent collection of Estonian First Encounters data. The MINT project aims at deepening our understanding of the conversational properties and practices in human interactions. We especially investigate conversational engagement and cooperation, and discuss some observations on the participants' views concerning the interaction they have been engaged.

DISLOG: A logic-based language for processing discourse structures

Patrick Saint-Dizier

In this paper, we present the foundations and the properties of the DISLOG language, a logic-based language designed to describe and implement discourse structure analysis. Dislog has the flexibility and the expressiveness of a rule-based system, it offers the possibility to include knowledge and reasoning capabilities and the expression a variety of well-formedness constraints proper to discourse. Dislog is embedded into the <TextCoop> platform that offers an engine with various processing capabilities and a programming environment.

A Repository of Rules and Lexical Resources for Discourse Structure Analysis: the Case of Explanation Structures

Sarah Bourse and Patrick Saint-Dizier

In this paper, we present an analysis method, a set of rules, lexical resources dedicated to discourse relation identification, in particular for explanation analysis. The following relations are described with prototypical rules: instructions, advice, warnings, illustration, restatement, purpose, condition, circumstance, concession, contrast and some forms of causes. Rules are developed for French and English. The approach used to describe the analysis of such relations is basically generative and also provides a conceptual view of explanation. The implementation is realized in Dislog, using the <TextCoop> logic-based platform, and the Dislog language, that also allows for the integration of knowledge and reasoning into rules describing the structure of explanation.

Feature Discovery for Diachronic Register Analysis: a Semi-Automatic Approach

Stefania Degaetano-Ortlieb, Ekaterina Lapshinova-Koltunski and Elke Teich

In this paper, we present corpus-based procedures to semi-automatically discover features relevant for the study of recent language change in scientific registers. First, linguistic features potentially adherent to recent language change are extracted from the SciTex Corpus. Second, features are assessed for their relevance for the study of recent language change in scientific registers by means of correspondence analysis. The discovered features will serve for further investigations of the linguistic evolution of newly emerged scientific registers.

Improving the Recall of a Discourse Parser by Constraint-based Postprocessing

Sucheta Ghosh, Richard Johansson, Giuseppe Riccardi and Sara Tonelli

We describe two constraint-based methods that can be used to improve the recall of a shallow discourse parser based on conditional random field chunking. These method uses a set of natural structural constraints as well as others that follow from the annotation guidelines of the Penn Discourse Treebank. We evaluated the resulting systems on the standard test set of the PDTB and achieved a rebalancing of precision and recall with improved F-measures across the board. This was especially notable when we used evaluation metrics taking partial matches into account; for these measures, we achieved F-measure improvements of several points.

Annotating dropped pronouns in Chinese newswire text

Elizabeth Baran, Yaqin Yang and Nianwen Xue

We propose an annotation framework to explicitly identify dropped subject pronouns in Chinese. We acknowledge and specify 10 concrete pronouns that exist as words in Chinese and 4 abstract pronouns that do not correspond to Chinese words, but that are recognized conceptually, to native Chinese speakers. These abstract pronouns are identified as “unspecified", “pleonastic", “event", and “existential” and are argued to exist cross-linguistically. We trained two annotators, fluent in Chinese, and adjudicated their annotations to form a gold standard. We achieved an inter-annotator agreement kappa of .6 and an observed agreement of .7. We found that annotators had the most difficulty with the abstract pronouns, such as “unspecified” and “event", but we posit that further specification and training has the potential to significantly improve these results. We believe that this annotated data will serve to help improve Machine Translation models that translate from Chinese to a non pro-drop language, like English, that requires all subject pronouns to be explicit.

Alternative Lexicalizations of Discourse Connectives in Czech

Magdalena Rysova

The paper concentrates on which language means may be included into the annotation of discourse relations in the Prague Dependency Treebank (PDT) and tries to examine the so called alternative lexicalizations of discourse markers (AltLex’s) in Czech. The analysis proceeds from the annotated data of PDT and tries to draw a comparison between the Czech AltLex’s from PDT and English AltLex’s from PDTB (the Penn Discourse Treebank). The paper presents a lexico-syntactic and semantic characterization of the Czech AltLex’s and comments on the current stage of their annotation in PDT. In the current version, PDT contains 306 expressions (within the total 43,955 of sentences) that were labeled by annotators as being an AltLex. However, as the analysis demonstrates, this number is not final. We suppose that it will increase after the further elaboration, as

AltLex's are not restricted to a limited set of syntactic classes and some of them exhibit a great degree of variation.

METU Turkish Discourse Bank Browser

Utku Şirin, Ruket Çakıcı and Deniz Zeyrek

In this paper, the METU Turkish Discourse Bank Browser, a tool developed for browsing the annotated annotated discourse relations in Middle East Technical University (METU) Turkish Discourse Bank (TDB) project is presented. The tool provides both a clear interface for browsing the annotated corpus and a wide range of search options to analyze the annotations.

DramaBank: Annotating Agency in Narrative Discourse

David Elson

We describe the Story Intention Graph, a set of discourse relations designed to represent aspects of narrative. Compared to prior models, ours is a novel synthesis of the notions of goal, plan, intention, outcome, affect and time that is amenable to corpus annotation. We describe a collection project, DramaBank, which includes encodings of texts ranging from small fables to epic poetry and contemporary nonfiction.

Multi-Layer Discourse Annotation of a Dutch Text Corpus

Gisela Redeker, Ildikó Berzlánovich, Nynke van der Vliet, Gosse Bouma and Markus Egg

We have compiled a corpus of 80 Dutch texts from expository and persuasive genres, which we annotated for rhetorical and genre-specific discourse structure, and lexical cohesion with the goal of creating a gold standard for further research. The annotations are based on a segmentation of the text in elementary discourse units that takes into account cues from syntax and punctuation. During the labor-intensive discourse-structure annotation (RST analysis), we took great care to thoroughly reconcile the initial analyses. That process and the availability of two independent initial analyses for each text allows us to analyze our disagreements and to assess the confusability of RST relations, and thereby improve the annotation guidelines and gather evidence for the classification of these relations into larger groups. We are using this resource for corpus-based studies of discourse relations, discourse markers, cohesion, and genre differences, e.g., the question of how discourse structure and lexical cohesion interact for different genres in the overall organization of texts. We are also exploring automatic text segmentation and semi-automatic discourse annotation.

Clause-based Discourse Segmentation of Arabic Texts

Iskandar Keskes, Farah Benamara and Lamia Hadrich Belguith

This paper describes a rule-based approach to segment Arabic texts into clauses. Our method relies on an extensive analysis of a large set of lexical cues as well as punctuation marks. Our analysis was carried out on two different corpus genres: news articles and elementary school textbooks. We propose a three steps segmentation algorithm: first by using only punctuation marks, then by relying only on lexical cues and finally by using both typology and lexical cues. The results were compared with manual segmentations elaborated by experts.

Project FLY: a multidisciplinary project within Linguistics

Mariana Gomes, Ana Guilherme, Leonor Tavares and Rita Marquilhas

This paper concerns the presentation of two projects that aim to make available an online archive of 4,000 original private letters, mainly having in mind research in Linguistics (Corpus Linguistics, Historical Linguistics, Pragmatics, Sociolinguistics, General Linguistics), History and Sociology. Our corpus is prepared for each research area and provides a diachronic archive of the Portuguese language. Projects CARDS and FLY have the main goal of making available an online electronic edition of each letter, which is completely open source, searchable and available. Users can search for an individual letter, a text by type, a group of letters by year or even the whole archive as a corpus for research or other purposes. The means of corpus presentation is a multimodal framework, since it joins together both the manuscript's image and the written text: the letter's material representation in facsimile and the letter's digital transcription. This editing method allows for the possibility of creating an annotated corpus where the textual unity is not lost.

Revealing Contentious Concepts Across Social Groups

Ching-Sheng Lin, Zumrut Akcam, Samira Shaikh, Sharon Small, Ken Stahl, Tomek Strzalkowski and Nick Webb

In this paper, a computational model based on concept polarity is proposed to investigate the influence of communications across the diacultural groups. The hypothesis of this work is that there are communities or groups which can be characterized by a network of concepts and the corresponding valuations of those concepts

that are agreed upon by the members of the community. We apply an existing research tool, ECO, to generate text representative of each community and create community specific Valuation Concept Networks (VCN). We then compare VCNs across the communities, to attempt to find contentious concepts, which could subsequently be the focus of further exploration as points of contention between the two communities. A prototype, CPAM (Changing Positions, Altering Minds), was implemented as a proof of concept for this approach. The experiment was conducted using blog data from pro-Palestinian and pro-Israeli communities. A potential application of this method and future work are discussed as well.

P31 - Lexical Acquisition

Thursday, May 24, 18:20

Chairperson: **Pierre Zweigenbaum** Poster Session

Customizable SCF Acquisition in Italian

Tommaso Caselli, Francesco Rubino, Francesca Frontini, Irene Russo and Valeria Quochi

Lexica of predicate-argument structures constitute a useful tool for several tasks in NLP. This paper describes a web-service system for automatic acquisition of verb subcategorization frames (SCFs) from parsed data in Italian. The system acquires SCFs in an unsupervised manner. We created two gold standards for the evaluation of the system, the first by mixing together information from two lexica (one manually created and the second automatically acquired) and manual exploration of corpus data and the other annotating data extracted from a specialized corpus (environmental domain). Data filtering is accomplished by means of the maximum likelihood estimate (MLE). The evaluation phase has allowed us to identify the best empirical MLE threshold for the creation of a lexicon (P=0.653, R=0.557, F1=0.601). In addition to this, we assigned to the extracted entries of the lexicon a confidence score based on the relative frequency and evaluated the extractor on domain specific data. The confidence score will allow the final user to easily select the entries of the lexicon in terms of their reliability: one of the most interesting feature of this work is the possibility the final users have to customize the results of the SCF extractor, obtaining different SCF lexica in terms of size and accuracy.

Large Scale Lexical Analysis

Gregor Thurmair, Vera Aleksic and Christoph Schwarz

The following paper presents a lexical analysis component as implemented in the PANACEA project. The goal is to automatically extract lexicon entries from crawled corpora, in an attempt to use corpus-based methods for high-quality linguistic text processing, and to focus on the quality of data without neglecting quantitative aspects. Lexical analysis has the task to assign linguistic information (like: part of speech, inflectional class, gender, subcategorisation frame, semantic properties etc.) to all parts of the input text. If tokens are ambiguous, lexical analysis must provide all possible sets of annotation for later (syntactic) disambiguation, be it tagging, or full parsing. The paper presents an approach for assigning part-of-speech tags for German and English to large input corpora (> 50 mio tokens), providing a workflow which takes as input crawled corpora and provides POS-tagged lemmata ready for lexicon integration. Tools include sentence splitting, lexicon lookup, decomposition, and POS defaulting. Evaluation shows that the overall error rate can be brought down to about 2% if language resources are properly designed. The complete workflow is implemented as a sequence of web services integrated into the PANACEA platform.

Extending the adverbial coverage of a French morphological lexicon

Elsa Tolone, Stavroula Voyatzi, Claude Martineau and Matthieu Constant

We present an extension of the adverbial entries of the French morphological lexicon DELA (Dictionnaires Electroniques du LADL / LADL electronic dictionaries). Adverbs were extracted from LGLex, a NLP-oriented syntactic resource for French, which in its turn contains all adverbs extracted from the Lexicon-Grammar tables of both simple adverbs ending in -ment (i.e., '-ly') and compound adverbs. This work exploits fine-grained linguistic information provided in existing resources. The resulting resource is reviewed in order to delete duplicates and is freely available under the LGPL-LR license.

Corpus based Semi-Automatic Extraction of Persian Compound Verbs and their Relations

Somayeh Bagherbeygi and Mehrnoush Shamsfard

Nowadays, Wordnet is used in natural language processing as one of the major linguistic resources. Having such a resource for Persian language helps researchers in computational linguistics and natural language processing fields to develop more accurate systems with higher performances. In this research, we propose a model for semi-automatic construction of Persian wordnet of verbs. Compound verbs are a very productive structure in Persian and number of compound verbs is much greater than simple verbs in this language This research is aimed at finding the structure of Persian compound verbs and the relations between verb components. The main idea behind developing this system is using the wordnet of other POS categories (here

means noun and adjective) to extract Persian compound verbs, their synsets and their relations. This paper focuses on three main tasks: 1.extracting compound verbs 2.extracting verbal synsets and 3.extracting the relations among verbal synsets such as hypernymy, antonymy and cause.

P32 - Corpus Creation, Processing, Usage (2)

Thursday, May 24, 18:20

Chairperson: **Shyam Agrawal** Poster Session

Extending the MPC corpus to Chinese and Urdu - A Multiparty Multi-Lingual Chat Corpus for Modeling Social Phenomena in Language

Ting Liu, Samira Shaikh, Tomek Strzalkowski, Aaron Broadwell, Jennifer Stromer-Galley, Sarah Taylor, Umit Boz, Xiaoai Ren and Jingsi Wu

In this paper, we report our efforts in building a multi-lingual multi-party online chat corpus in order to develop a firm understanding in a set of social constructs such as agenda control, influence, and leadership as well as to computationally model such constructs in online interactions. These automated models will help capture the dialogue dynamics that are essential for developing, among others, realistic human-machine dialogue systems, including autonomous virtual chat agents. In this paper, we first introduce our experiment design and data collection method in Chinese and Urdu, and then report on the current stage of our data collection. We annotated the collected corpus on four levels: communication links, dialogue acts, local topics, and meso-topics. Results from the analyses of annotated data on different languages indicate some interesting phenomena, which are reported in this paper.

Multimedia database of the cultural heritage of the Balkans

Ivana Tanasijević, Biljana Sikimić and Gordana Pavlović-Lažetić

This paper presents a system that is designed to make possible the organization and search within the collected digitized material of intangible cultural heritage. The motivation for building the system was a vast quantity of multimedia documents collected by a team from the Institute for Balkan Studies in Belgrade. The main topic of their research were linguistic properties of speeches that are used in various places in the Balkans by different groups of people. This paper deals with a prototype system that enables the annotation of the collected material and its organization into a native XML database through a graphical interface. The system enables the search of the database and the presentation of digitized multimedia documents and spatial as well as non-spatial information of the queried data. The multimedia content can be read, listened to or watched while spatial properties are presented on the graphics that consists of geographic regions in the Balkans. The system also enables spatial queries by consulting the graph of geographic regions.

YADAC: Yet another Dialectal Arabic Corpus

Rania Al-Sabbagh and Roxana Girju

This paper presents the first phase of building YADAC – a multi-genre Dialectal Arabic (DA) corpus – that is compiled using Web data from microblogs (i.e. Twitter), blogs/forums and online knowledge market services in which both questions and answers are user-generated. In addition to introducing two new genres to the current efforts of building DA corpora (i.e. microblogs and question-answer pairs extracted from online knowledge market services), the paper highlights and tackles several new issues related to building DA corpora that have not been handled in previous studies: function-based Web harvesting and dialect identification, vowel-based spelling variation, linguistic hypercorrection and its effect on spelling variation, unsupervised Part-of-Speech (POS) tagging and base phrase chunking for DA. Although the algorithms for both POS tagging and base-phrase chunking are still under development, the results are promising.

The Trilingual ALLEGRA Corpus: Presentation and Possible Use for Lexicon Induction

Yves Scherrer and Bruno Cartoni

In this paper, we present a trilingual parallel corpus for German, Italian and Romansh, a Swiss minority language spoken in the canton of Grisons. The corpus called ALLEGRA contains press releases automatically gathered from the website of the cantonal administration of Grisons. Texts have been preprocessed and aligned with a current state-of-the-art sentence aligner. The corpus is one of the first of its kind, and can be of great interest, particularly for the creation of natural language processing resources and tools for Romansh. We illustrate the use of such a trilingual resource for automatic induction of bilingual lexicons, which is a real challenge for under-represented languages. We induce a bilingual lexicon for German-Romansh by phrase alignment and evaluate the resulting entries with the help of a reference lexicon. We then show that the use of the third language of the corpus – Italian – as a pivot language can improve the precision of the induced lexicon, without loss in terms of quality of the extracted pairs.

Beyond SoNaR: towards the facilitation of large corpus building efforts

Martin Reynaert, Ineke Schuurman, Veronique Hoste, Nelleke Oostdijk and Maarten van Gompel

In this paper we report on the experiences gained in the recent construction of the SoNaR corpus, a 500 MW reference corpus of contemporary, written Dutch. It shows what can realistically be done within the confines of a project setting where there are limitations to the duration in time as well to the budget, employing current state-of-the-art tools, standards and best practices. By doing so we aim to pass on insights that may be beneficial for anyone considering to undertake an effort towards building a large, varied yet balanced corpus for use by the wider research community. Various issues are discussed that come into play while compiling a large corpus, including approaches to acquiring texts, the arrangement of IPR, the choice of text formats, and steps to be taken in the preprocessing of data from widely different origins. We describe FoLiA, a new XML format geared at rich linguistic annotations. We also explain the rationale behind the investment in the high-quali ty semi-automatic enrichment of a relatively small (1 MW) subset with very rich syntactic and semantic annotations. Finally, we present some ideas about future developments and the direction corpus development may take, such as setting up an integrated work flow between web services and the potential role for ISOcat. We list tips for potential corpus builders, tricks they may want to try and further recommendations regarding technical developments future corpus builders may wish to hope for.

The New IDS Corpus Analysis Platform: Challenges and Prospects

Piotr Bański, Peter M. Fischer, Elena Frick, Erik Ketzan, Marc Kupietz, Carsten Schnober, Oliver Schonefeld and Andreas Witt

The present article describes the first stage of the KorAP project, launched recently at the Institut für Deutsche Sprache (IDS) in Mannheim, Germany. The aim of this project is to develop an innovative corpus analysis platform to tackle the increasing demands of modern linguistic research. The platform will facilitate new linguistic findings by making it possible to manage and analyse primary data and annotations in the petabyte range, while at the same time allowing an undistorted view of the primary linguistic data, and thus fully satisfying the demands of a scientific tool. An additional important aim of the project is to make corpus data as openly accessible as possible in light of unavoidable legal restrictions, for instance through support for distributed virtual corpora, user-defined annotations and adaptable user interfaces, as well as interfaces and sandboxes for user-supplied analysis applications. We discuss our motivation for undertaking this endeavour and the challenges that face it. Next, we outline our software implementation plan and describe development to-date.

A Tool/Database Interface for Multi-Level Analyses

Kurt Eberle, Kerstin Eckart, Ulrich Heid and Boris Haselbach

Depending on the nature of a linguistic theory, empirical investigations of its soundness may focus on corpus studies related to lexical, syntactic, semantic or other phenomena. Especially work in research networks usually comprises analyses of different levels of description, where each one must be as reliable as possible when the same sentences and texts are investigated under very different perspectives. This paper describes an infrastructure that interfaces an analysis tool for multi-level annotation with a generic relational database. It supports three dimensions of analysis-handling and thereby builds an integrated environment for quality assurance in corpus based linguistic analysis: a vertical dimension relating analysis components in a pipeline, a horizontal dimension taking alternative results of the same analysis level into account and a temporal dimension to follow up cases where analyses for the same input have been produced with different versions of a tool. As an example we give a detailed description of a typical workflow for the vertical dimension.

New language resources for the Pashto language

Djamel Mostefa, Khalid Choukri, Sylvie Brunessaux, Karim Boudahmane

This paper reports on the development of new language resources for the Pashto language, a very low-resource language spoken in Afghanistan and Pakistan. In the scope of a multilingual data collection project, three large corpora are collected for Pashto. Firstly a monolingual text corpus of 100 million words is produced. Secondly a 100 hours speech database is recorded and manually transcribed. Finally a bilingual Pashto-French parallel corpus of around 2 million is produced by translating Pashto texts into French. These resources will be used to develop Human Language Technology systems for Pashto with a special focus on Machine Translation.

CALBC: Releasing the Final Corpora

Şenay Kafkas, Ian Lewin, David Milward, Erik van Mulligen, Jan Kors, Udo Hahn and Dietrich Rebholz-Schuhmann

A number of gold standard corpora for named entity recognition are available to the public. However, the existing gold standard

corpora are limited in size and semantic entity types. These usually lead to implementation of trained solutions (1) for a limited number of semantic entity types and (2) lacking in generalization capability. In order to overcome these problems, the CALBC project has aimed to automatically generate large scale corpora annotated with multiple semantic entity types in a community-wide manner based on the consensus of different named entity solutions. The generated corpus is called the silver standard corpus since the corpus generation process does not involve any manual curation. In this publication, we announce the release of the final CALBC corpora which include the silver standard corpus in different versions and several gold standard corpora for the further usage of the biomedical text mining community. The gold standard corpora are utilised to benchmark the methods used in the silver standard corpora generation process and released in a shared format. All the corpora are released in a shared format and accessible at www.calbc.eu.

Language Richness of the Web

Martin Majliš and Zdeněk Žabokrtský

We have built a corpus containing texts in 106 languages from texts available on the Internet and on Wikipedia. The W2C Web Corpus contains 54.7 GB of text and the W2C Wiki Corpus contains 8.5 GB of text. The W2C Web Corpus contains more than 100 MB of text available for 75 languages. At least 10 MB of text is available for 100 languages. These corpora are a unique data source for linguists, since they outclass all published works both in the size of the material collected and the number of languages covered. This language data resource can be of use particularly to researchers specialized in multilingual technologies development. We also developed software that greatly simplifies the creation of a new text corpus for a given language, using text materials freely available on the Internet. Special attention was given to components for filtering and de-duplication that allow to keep the material quality very high.

P33 - Web Services

Thursday, May 24, 18:20

Chairperson: **Nuria Bel** Poster Session

Cloud Logic Programming for Integrating Language Technology Resources

Markus Forsberg and Torbjörn Lager

The main goal of the CLT Cloud project is to equip lexica, morphological processors, parsers and other software components developed within CLT (Centre of Language Technology) with so called web API:s, thus making them available on the Internet in the form of web services. We present a proof-of-concept implementation of the CLT Cloud server where we use the logic programming language Prolog for composing and aggregating existing web services into new web services in a way that encourages creative exploration and rapid prototyping of LT applications.

Dynamic web service deployment in a cloud environment

Marc Kemps-Snijders, Matthijs Brouwer, Jan Pieter Kunst and Tom Visser

E-infrastructure projects such as CLARIN do not only make research data available to the scientific community, but also deliver a growing number of web services. While the standard methods for deploying web services using dedicated (virtual) server may suffice in many circumstances, CLARIN centers are also faced with a growing number of services that are not frequently used and for which significant compute power needs to be reserved. This paper describes an alternative approach towards service deployment capable of delivering on demand services in a workflow using cloud infrastructure capabilities. Services are stored as disk images and deployed on a workflow scenario only when needed this helping to reduce the overall service footprint.

Word Sketches for Turkish

Bharat Ram Ambati, Siva Reddy and Adam Kilgarriff

Word sketches are one-page, automatic, corpus-based summaries of a word's grammatical and collocational behaviour. In this paper we present word sketches for Turkish. Until now, word sketches have been generated using a purpose-built finite-state grammars. Here, we use an existing dependency parser. We describe the process of collecting a 42 million word corpus, parsing it, and generating word sketches from it. We evaluate the word sketches in comparison with word sketches from a language independent sketch grammar on an external evaluation task called topic coherence, using Turkish WordNet to derive an evaluation set of coherent topics.

Service Composition Scenarios for Task-Oriented Translation

Chunqi Shi, Donghui Lin and Toru Ishida

Due to instant availability and low cost, machine translation is becoming popular. Machine translation mediated communication plays a more and more important role in international collaboration. However, machine translators cannot guarantee high quality translation. In a multilingual communication task, many in-domain resources, for example domain dictionaries, are needed to promote translation quality. This raises the problem of

how to help communication task designers provide higher quality translation systems, systems that can take advantage of various in-domain resources. The Language Grid, a service-oriented collective intelligent platform, allows in-domain resources to be wrapped into language services. For task-oriented translation, we propose service composition scenarios for the composition of different language services, where various in-domain resources are utilized effectively. We design the architecture, provide a script language as the interface for the task designer, which is easy for describing the composition scenario, and make a case study of a Japanese-English campus orientation task. Based on the case study, we analyze the increase in translation quality possible and the usage of in-domain resources. The results demonstrate a clear improvement in translation accuracy when the in-domain resources are used.

Linguistic Analysis Processing Line for Bulgarian

Aleksandar Savkov, Laska Laskova, Stanislava Kancheva, Petya Osenova and Kiril Simov

This paper presents a linguistic processing pipeline for Bulgarian including morphological analysis, lemmatization and syntactic analysis of Bulgarian texts. The morphological analysis is performed by three modules – two statistical-based and one rule-based. The combination of these modules achieves the best result for morphological tagging of Bulgarian over a rich tagset (680 tags). The lemmatization is based on rules, generated from a large morphological lexicon of Bulgarian. The syntactic analysis is implemented via MaltParser. The two statistical morphological taggers and MaltParser are trained on datasets constructed within BulTreeBank project. The processing pipeline includes also a sentence splitter and a tokenizer. All tools in the pipeline are packed in modules that can also perform separately. The whole pipeline is designed to be able to serve as a back-end of a web service oriented interface, but it also supports the user tasks with a command-line interface. The processing pipeline is compatible with the Text Corpus Format, which allows it to delegate the management of the components to the WebLicht platform.

On the Way to a Legal Sharing of Web Applications in NLP

Victoria Arranz and Olivier Hamon

For some years now, web services have been employed in Natural Language Processing (NLP) for a number of uses and within a number of sub-areas. Web services allow users to gain access to distant applications without having the need to install them on their local machines. A large paradigm of advantages can be obtained from a practical and development point of view. However, the legal aspects behind this sharing should not be neglected and should be openly discussed so as to understand the implications behind such data exchanges and tool uses. In the framework of PANACEA, this paper highlights the different points involved and describes the work done in order to handle all the legal aspects behind those points.

Collaborative Development and Evaluation of Text-processing Workflows in a UIMA-supported Web-based Workbench

Rafal Rak, Andrew Rowley and Sophia Ananiadou

Challenges in creating comprehensive text-processing worklows include a lack of the interoperability of individual components coming from different providers and/or a requirement imposed on the end users to know programming techniques to compose such workflows. In this paper we demonstrate Argo, a web-based system that addresses these issues in several ways. It supports the widely adopted Unstructured Information Management Architecture (UIMA), which handles the problem of interoperability; it provides a web browser-based interface for developing workflows by drawing diagrams composed of a selection of available processing components; and it provides novel user-interactive analytics such as the annotation editor which constitutes a bridge between automatic processing and manual correction. These features extend the target audience of Argo to users with a limited or no technical background. Here, we focus specifically on the construction of advanced workflows, involving multiple branching and merging points, to facilitate various comparative evalutions. Together with the use of user-collaboration capabilities supported in Argo, we demonstrate several use cases including visual inspections, comparisions of multiple processing segments or complete solutions against a reference standard, inter-annotator agreement, and shared task mass evaluations. Ultimetely, Argo emerges as a one-stop workbench for defining, processing, editing and evaluating text processing tasks.

The SERENOA Project: Multidimensional Context-Aware Adaptation of Service Front-Ends

Javier Caminero, Mari Carmen Rodríguez, Jean Vanderdonckt, Fabio Paternò, Joerg Rett, Dave Raggett, Jean-Loup Comeliau and Ignacio Marín

The SERENOA project is aimed at developing a novel, open platform for enabling the creation of context-sensitive Service Front-Ends (SFEs). A context-sensitive SFE provides a user interface (UI) that allows users to interact with remote services, and which exhibits some capability to be aware of the context and to react to changes of this context in a continuous way. As a result, such UI will be adapted to e.g. a person's devices,

tasks, preferences, abilities, and social relationships, as well as the conditions of the surrounding physical environment, thus improving people's satisfaction and performance compared to traditional SFEs based on manually designed UIs. The final aim is to support humans in a more effective, personalized and consistent way, thus improving the quality of life for citizens. In this scenario, we envisage SERENOA as the reference implementation of a SFE adaptation platform for the 'Future Internet'.

Keynote Speech 2

Friday, May 25, 9:00

Chairperson: **Asuncion Moreno**

Socio-emotional virtual agent

Catherine Pelachaud

In this talk I will present our current work toward endowing virtual agents with communicative and emotional capabilities. I will start describing an interactive system of an agent dialoging with human users in an emotionally colored manner. The agent is an active listener and can display a large range of back-channels. Four agents have been defined with specific emotionally traits. They are defined by specific facial models but also behaviors. They can also mimic user's expressions such as smiles and head movements. Through their behaviors, the agents show various attitudes and levels of engagement. Lately we have concentrated our study on smile. Smiles can have many communicative functions, such as being a smile of politeness, happiness or embarrassment. These smiles differ in their morphology, being of shorter or longer duration, symmetric or not. Agents are perceived with different attitudes when they display these smiles.

O37 - Subjectivity and Emotions

Friday, May 25, 9:45

Chairperson: **Julio Gonzalo** Oral Session

Towards Emotion and Affect Detection in the Multimodal LAST MINUTE Corpus

Jörg Frommer, Bernd Michaelis, Dietmar Rösner, Andreas Wendemuth, Rafael Friesen, Matthias Haase, Manuela Kunze, Rico Andrich, Julia Lange, Axel Panning and Ingo Siegert

The LAST MINUTE corpus comprises multimodal recordings (e.g. video, audio, transcripts) from WOZ interactions in a mundane planning task (Rösner et al., 2011). It is one of the largest corpora with naturalistic data currently available. In this paper we report about first results from attempts to automatically and manually analyze the different modes with respect to emotions and affects exhibited by the subjects. We describe and discuss difficulties encountered due to the strong contrast between the naturalistic recordings and traditional databases with acted emotions.

Building a fine-grained subjectivity lexicon from a web corpus

Isa Maks and Piek Vossen

In this paper we propose a method to build fine-grained subjectivity lexicons including nouns, verbs and adjectives. The method, which is applied for Dutch, is based on the comparison of word frequencies of three corpora: Wikipedia, News and News comments. Comparison of the corpora is carried out with two measures: log-likelihood ratio and a percentage difference calculation. The first step of the method involves subjectivity identification, i.e. determining if a word is subjective or not. The second step aims at the identification of more fine-grained subjectivity which is the distinction between actor subjectivity and speaker / writer subjectivity. The results suggest that this approach can be usefully applied producing subjectivity lexicons of high quality.

Learning Sentiment Lexicons in Spanish

Veronica Perez-Rosas, Carmen Banea and Rada Mihalcea

In this paper we present a framework to derive sentiment lexicons in a target language by using manually or automatically annotated data available in an electronic resource rich language, such as English. We show that bridging the language gap using the multilingual sense-level aligned WordNet structure allows us to generate a high accuracy (90%) polarity lexicon comprising 1,347 entries, and a disjoint lower accuracy (74%) one encompassing 2,496 words. By using an LSA-based vectorial expansion for the generated lexicons, we are able to obtain an average F-measure of 66% in the target language. This implies that the lexicons could be used to bootstrap higher-coverage lexicons using in-language resources.

Assigning Connotation Values to Events

Tommaso Caselli, Irene Russo and Francesco Rubino

Sentiment Analysis (SA) and Opinion Mining (OM) have become a popular task in recent years in NLP with the development of language resources, corpora and annotation schemes. The possibility to discriminate between objective and subjective expressions contributes to the identification of a document's semantic orientation and to the detection of the opinions and sentiments expressed by the authors or attributed

to other participants in the document. Subjectivity word sense disambiguation helps in this task, automatically determining which word senses in a corpus are being used subjectively and which are being used objectively. This paper reports on a methodology to assign in a semi-automatic way connotative values to eventive nouns usually labelled as neutral through syntagmatic patterns that express cause-effect relations between emotion cause events and emotion words. We have applied our method to nouns and we have been able reduce the number of OBJ polarity values associated to event noun.

Cost and Benefit of Using WordNet Senses for Sentiment Analysis

Balamuraliar, Aditya Joshi and Pushpak Bhattacharyya

Typically, accuracy is used to represent the performance of an NLP system. However, accuracy attainment is a function of investment in annotation. Typically, the more the amount and sophistication of annotation, higher is the accuracy. However, a moot question is“is the accuracy improvement commensurate with the cost incurred in annotation"? We present an economic model to assess the marginal benefit accruing from increase in cost of annotation. In particular, as a case in point we have chosen the sentiment analysis (SA) problem. In SA, documents normally are polarity classified by running them through classifiers trained on document vectors constructed from lexeme features, i.e., words. If, however, instead of words, one uses word senses (synset ids in wordnets) as features, the accuracy improves dramatically. But is this improvement significant enough to justify the cost of annotation? This question, to the best of our knowledge, has not been investigated with the seriousness it deserves. We perform a cost benefit study based on a vendor-machine model. By setting up a cost price, selling price and profit scenario, we show that although extra cost is incurred in sense annotation, the profit margin is high, justifying the cost.

O38 - Named Entities

Friday, May 25, 9:45

Chairperson: **Satoshi Sato** Oral Session

Linguistic Resources for Entity Linking Evaluation: from Monolingual to Cross-lingual

Xuansong Li, Stephanie Strassel, Heng Ji, Kira Griffitt and Joe Ellis

To advance information extraction and question answering technologies toward a more realistic path, the U.S. NIST (National Institute of Standards and Technology) initiated the KBP (Knowledge Base Population) task as one of the TAC (Text Analysis Conference) evaluation tracks. It aims to encourage research in automatic information extraction of named entities from unstructured texts with the ultimate goal of integrating such information into a structured Knowledge Base. The KBP track consists of two types of evaluation: Named Entity Linking (NEL) and Slot Filling. This paper describes the linguistic resource creation efforts at the Linguistic Data Consortium (LDC) in support of Named Entity Linking evaluation of KBP, focusing on annotation methodologies, process, and features of corpora from 2009 to 2011, with a highlighted analysis of the cross-lingual NEL data. Progressing from monolingual to cross-lingual Entity Linking technologies, the 2011 cross-lingual NEL evaluation targeted multilingual capabilities. Annotation accuracy is presented in comparison with system performance, with promising results from cross-lingual entity linking systems.

Creating and Curating a Cross-Language Person-Entity Linking Collection

Dawn Lawrie, James Mayfield, Paul McNamee and Douglas Oard

To stimulate research in cross-language entity linking, we present a new test collection for evaluating the accuracy of cross-language entity linking in twenty-one languages. This paper describes an efficient way to create and curate such a collection, judiciously exploiting existing language resources. Queries are created by semi-automatically identifying person names on the English side of a parallel corpus, using judgments obtained through crowdsourcing to identify the entity corresponding to the name, and projecting the English name onto the non-English document using word alignments. Name projections are then curated, again through crowdsourcing. This technique resulted in the first publicly available multilingual cross-language entity linking collection. The collection includes approximately 55,000 queries, comprising between 875 and 4,329 queries for each of twenty-one non-English languages.

International Multicultural Name Matching Competition: Design, Execution, Results, and Lessons Learned

Keith J. Miller, Elizabeth Schroeder Richerson, Sarah McLeod, James Finley and Aaron Schein

This paper describes different aspects of an open competition to evaluate multicultural name matching software, including the contest design, development of the test data, different phases of the competition, behavior of the participating teams, results of the competition, and lessons learned throughout. The competition, known as The MITRE Challenge™, was informally announced at LREC 2010 and was recently concluded. Contest participants

used the competition website (http://mitrechallenge.mitre.org) to download the competition data set and guidelines, upload results, and to view accuracy metrics for each result set submitted. Participants were allowed to submit unlimited result sets, with their top-scoring set determining their overall ranking. The competition website featured a leader board that displayed the top score for each participant, ranked according to the principal contest metric - mean average precision (MAP). MAP and other metrics were calculated in near-real time on a remote server, based on ground truth developed for the competition data set. Additional measures were taken to guard against gaming the competition metric or overfitting to the competition data set. Lessons learned during running this first MITRE Challenge will be valuable to others considering running similar evaluation campaigns.

An Empirical Study of the Occurrence and Co-Occurrence of Named Entities in Natural Language Corpora

K Saravanan, Monojit Choudhury, Raghavendra Udupa and A Kumaran

Named Entities (NEs) that occur in natural language text are important especially due to the advent of social media, and they play a critical role in the development of many natural language technologies. In this paper, we systematically analyze the patterns of occurrence and co-occurrence of NEs in standard large English news corpora - providing valuable insight for the understanding of the corpus, and subsequently paving way for the development of technologies that rely critically on handling NEs. We use two distinctive approaches: normal statistical analysis that measure and report the occurrence patterns of NEs in terms of frequency, growth, etc., and a complex networks based analysis that measures the co-occurrence pattern in terms of connectivity, degree-distribution, small-world phenomenon, etc. Our analysis indicates that: (i) NEs form an open-set in corpora and grow linearly, (ii) presence of a kernel and peripheral NE's, with the large periphery occurring rarely, and (iii) a strong evidence of small-world phenomenon. Our findings may suggest effective ways for construction of NE lexicons to aid efficient development of several natural language technologies.

Extended Named Entities Annotation on OCRed Documents: From Corpus Constitution to Evaluation Campaign

Olivier Galibert, Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum and Ludovic Quintard

Within the framework of the Quaero project, we proposed a new definition of named entities, based upon an extension of the coverage of named entities as well as the structure of those named entities. In this new definition, the extended named entities we proposed are both hierarchical and compositional. In this paper, we focused on the annotation of a corpus composed of press archives, OCRed from French newspapers of December 1890. We present the methodology we used to produce the corpus and the characteristics of the corpus in terms of named entities annotation. This annotated corpus has been used in an evaluation campaign. We present this evaluation, the metrics we used and the results obtained by the participants.

O39 - Treebanks and Syntax

Friday, May 25, 9:45

Chairperson: **Erhard Hinrichs** Oral Session

Making Ellipses Explicit in Dependency Conversion for a German Treebank

Wolfgang Seeker and Jonas Kuhn

We present a carefully designed dependency conversion of the German phrase-structure treebank TiGer that explicitly represents verb ellipses by introducing empty nodes into the tree. Although the conversion process uses heuristics like many other conversion tools we designed them to fail if no reasonable solution can be found. The failing of the conversion process makes it possible to detect elliptical constructions where the head is missing, but it also allows us to find errors in the original annotation. We discuss the conversion process and the heuristics, and describe some design decisions and error corrections that we applied to the corpus. Since most of today's data-driven dependency parsers are not able to handle empty nodes directly during parsing, our conversion tool also derives a canonical dependency format without empty nodes. It is shown experimentally to be well suited for training statistical dependency parsers by comparing the performance of two parsers from different parsing paradigms on the data set of the CoNLL 2009 Shared Task data and our corpus.

A Grammar-informed Corpus-based Sentence Database for Linguistic and Computational Studies

Hongzhi Xu, Helen Kaiyun Chen, Chu-Ren Huang, Qin Lu, Dingxu Shi and Tin-Shing Chiu

We adopt the corpus-informed approach to example sentence selections for the construction of a reference grammar. In the process, a database containing sentences that are carefully selected by linguistic experts including the full range of linguistic facts covered in an authoritative Chinese Reference Grammar is constructed and structured according to the reference grammar. A search engine system is developed to facilitate the process

of finding the most typical examples the users need to study a linguistic problem or prove their hypotheses. The database can also be used as a training corpus by computational linguists to train models for Chinese word segmentation, POS tagging and sentence parsing.

A Reference Dependency Bank for Analyzing Complex Predicates

Tafseer Ahmed, Miriam Butt, Annette Hautli and Sebastian Sulger

When dealing with languages of South Asia from an NLP perspective, a problem that repeatedly crops up is the treatment of complex predicates. This paper presents a first approach to the analysis of complex predicates (CPs) in the context of dependency bank development. The efforts originate in theoretical work on CPs done within Lexical-Functional Grammar (LFG), but are intended to provide a guideline for analyzing different types of CPs in an independent framework. Despite the fact that we focus on CPs in Hindi and Urdu, the design of the dependencies is kept general enough to account for CP constructions across languages.

Announcing Prague Czech-English Dependency Treebank 2.0

Jan Hajič, Eva Hajičová, Jarmila Panevová, Petr Sgall, Ondřej Bojar, Silvie Cinková, Eva Fučíková, Marie Mikulová, Petr Pajas, Jan Popelka, Jiří Semecký, Jana Šindlerová, Jan Štěpánek, Josef Toman, Zdeňka Urešová and Zdeněk Žabokrtský

We introduce a substantial update of the Prague Czech-English Dependency Treebank, a parallel corpus manually annotated at the deep syntactic layer of linguistic representation. The English part consists of the Wall Street Journal (WSJ) section of the Penn Treebank. The Czech part was translated from the English source sentence by sentence. This paper gives a high level overview of the underlying linguistic theory (the so-called tectogrammatical annotation) with some details of the most important features like valency annotation, ellipsis reconstruction or coreference.

Example-Based Treebank Querying

Liesbeth Augustinus, Vincent Vandeghinste and Frank Van Eynde

The recent construction of large linguistic treebanks for spoken and written Dutch (e.g. CGN, LASSY, Alpino) has created new and exciting opportunities for the empirical investigation of Dutch syntax and semantics. However, the exploitation of those treebanks requires knowledge of specific data structures and query languages such as XPath. Linguists who are unfamiliar with formal languages are often reluctant towards learning such a language. In order to make treebank querying more attractive for non-technical users we developed GrETEL (Greedy Extraction of Trees for Empirical Linguistics), a query engine in which linguists can use natural language examples as a starting point for searching the Lassy treebank without knowledge about tree representations nor formal query languages. By allowing linguists to search for similar constructions as the example they provide, we hope to bridge the gap between traditional and computational linguistics. Two case studies are conducted to provide a concrete demonstration of the tool. The architecture of the tool is optimised for searching the LASSY treebank, but the approach can be adapted to other treebank lay-outs.

O40 - Semantic Lexicons and Semantic Annotation

Friday, May 25, 9:45

Chairperson: **Dimitrios Kokkinakis** Oral Session

A Cross-Lingual Dictionary for English Wikipedia Concepts

Valentin I. Spitkovsky and Angel X. Chang

We present a resource for automatically associating strings of text with English Wikipedia concepts. Our machinery is bi-directional, in the sense that it uses the same fundamental probabilistic methods to map strings to empirical distributions over Wikipedia articles as it does to map article URLs to distributions over short, language-independent strings of natural language text. For maximal inter-operability, we release our resource as a set of flat line-based text files, lexicographically sorted and encoded with UTF-8. These files capture joint probability distributions underlying concepts (we use the terms article, concept and Wikipedia URL interchangeably) and associated snippets of text, as well as other features that can come in handy when working with Wikipedia articles and related information.

A database of semantic clusters of verb usages

Silvie Cinková, Martin Holub, Adam Rambousek and Lenka Smejkalová

We are presenting VPS-30-En, a small lexical resource that contains the following 30 English verbs: access, ally, arrive, breathe, claim, cool, crush, cry, deny, enlarge, enlist, forge, furnish, hail, halt, part, plough, plug, pour, say, smash, smell, steer, submit, swell, tell, throw, trouble, wake and yield. We have created and have been using VPS-30-En to explore the interannotator agreement potential of the Corpus Pattern Analysis. VPS-30-En is a small snapshot of the Pattern Dictionary

of English Verbs (Hanks and Pustejovsky, 2005), which we revised (both the entries and the annotated concordances) and enhanced with additional annotations. It is freely available at http://ufal.mff.cuni.cz/spr. In this paper, we compare the annotation scheme of VPS-30-En with the original PDEV. We also describe the adjustments we have made and their motivation, as well as the most pervasive causes of interannotator disagreements.

Is it Useful to Support Users with Lexical Resources? A User Study.

Ernesto William De Luca

Current search engines are used for retrieving relevant documents from the huge amount of data available and have become an essential tool for the majority of Web users. Standard search engines do not consider semantic information that can help in recognizing the relevance of a document with respect to the meaning of a query. In this paper, we present our system architecture and a first user study, where we show that the use of semantics can help users in finding relevant information, filtering it ad facilitating quicker access to data.

A review corpus annotated for negation, speculation and their scope

Natalia Konstantinova, Sheila C.M. de Sousa, Noa P. Cruz, Manuel J. Maña, Maite Taboada and Ruslan Mitkov

This paper presents a freely available resource for research on handling negation and speculation in review texts. The SFU Review Corpus, consisting of 400 documents of movie, book, and consumer product reviews, was annotated at the token level with negative and speculative keywords and at the sentence level with their linguistic scope. We report statistics on corpus size and the consistency of annotations. The annotated corpus will be useful in many applications, such as document mining and sentiment analysis.

Developing a large semantically annotated corpus

Valerio Basile, Johan Bos, Kilian Evang and Noortje Venhuizen

What would be a good method to provide a large collection of semantically annotated texts with formal, deep semantics rather than shallow? We argue that a bootstrapping approach comprising state-of-the-art NLP tools for parsing and semantic interpretation, in combination with a wiki-like interface for collaborative annotation of experts, and a game with a purpose for crowdsourcing, are the starting ingredients for fulfilling this enterprise. The result is a semantic resource that anyone can edit and that integrates various phenomena, including predicate-argument structure, scope, tense, thematic roles, rhetorical relations and presuppositions, into a single semantic formalism: Discourse Representation Theory. Taking texts rather than sentences as the units of annotation results in deep semantic representations that incorporate discourse structure and dependencies. To manage the various (possibly conflicting) annotations provided by experts and non-experts, we introduce a method that stores “Bits of Wisdom” in a database as stand-off annotations.

P34 Corpus Creation, Processing, Usage (3)

Friday, May 25, 9:45

Chairperson: **German Rigau** Poster Session

Le Petit Prince in UNL

Ronaldo Martins

The present paper addresses the process and the results of the interpretation of the integral text of “Le Petit Prince” (Little Prince), the famous novel by Antoine de Saint-Exupéry, from French into UNL. The original text comprised 1,684 interpretation units (15,513 words), which were sorted according to their similarity, from the shortest to the longest ones, and which were then projected into a UNL graph structure, composed of semantic directed binary relations linking nodes associated to the synsets of the corresponding original lexical items. The whole UNL-ization process was carried-out manually and the results have been used as the main resource in a natural language generation project involving already 27 languages.

A generic formalism to represent linguistic corpora in RDF and OWL/DL

Christian Chiarcos

This paper describes POWLA, a generic formalism to represent linguistic corpora by means of RDF and OWL/DL. Unlike earlier approaches in this direction, POWLA is not tied to a specific selection of annotation layers, but rather, it is designed to support any kind of text-oriented annotation. POWLA inherits its generic character from the underlying data model PAULA (Dipper, 2005; Chiarcos et al., 2009) that is based on early sketches of the ISO TC37/SC4 Linguistic Annotation Framework (Ide and Romary, 2004). As opposed to existing standoff XML linearizations for such generic data models, it uses RDF as representation formalism and OWL/DL for validation. The paper discusses advantages of this approach, in particular with respect to interoperability and queriability, which are illustrated for the MASC corpus, an open multi-layer corpus of American English (Ide et al., 2008).

A Database of Attribution Relations

Silvia Pareti

The importance of attribution is becoming evident due to its relevance in particular for Opinion Analysis and Information Extraction applications. Attribution would allow to identify different perspectives on a given topic or retrieve the statements of a specific source of interest, but also to select more relevant and reliable information. However, the scarce and partial resources available to date to conduct attribution studies have determined that only a portion of attribution structures has been identified and addressed. This paper presents the collection and further annotation of a database of over 9800 attributions relations from the Penn Discourse TreeBank (PDTB). The aim is to build a large and complete resource that fills a key gap in the field and enables the training and testing of robust attribution extraction systems.

KPWr: Towards a Free Corpus of Polish

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski and Adam Wardyński

This paper presents our efforts aimed at collecting and annotating a free Polish corpus. The corpus will serve for us as training and testing material for experiments with Machine Learning algorithms. As others may also benefit from the resource, we are going to release it under a Creative Commons licence, which is hoped to remove unnecessary usage restrictions, but also to facilitate reproduction of our experimental results. The corpus is being annotated with various types of linguistic entities: chunks and named entities, selected syntactic and semantic relations, word senses and anaphora. We report on the current state of the project as well as our ultimate goals.

Construction of the Turkish National Corpus (TNC)

Yeşim Aksan, Mustafa Aksan, Ahmet Koltuksuz, Taner Sezer, Ümit Mersinli, Umut Ufuk Demirhan, Hakan Yılmazer, Gülsüm Atasoy, Seda Öz, İpek Yıldız and Özlem Kurtoğlu

This paper addresses theoretical and practical issues experienced in the construction of Turkish National Corpus (TNC). TNC is designed to be a balanced, large scale (50 million words) and general-purpose corpus for contemporary Turkish. It has benefited from previous practices and efforts for the construction of corpora. In this sense, TNC generally follows the framework of British National Corpus, yet necessary adjustments in corpus design of TNC are made whenever needed. All throughout the process, different types of open-source software are used for specific tasks, and the resulting corpus is a free resource for non-commercial use. This paper presents TNC's design features, web-based corpus management system, carefully planned workflow and its web-based user-friendly search interface.

Building a learner corpus

Jirka Hana, Alexandr Rosen, Barbora Štindlová and Petr Jäger

The paper describes a corpus of texts produced by non-native speakers of Czech. We discuss its annotation scheme, consisting of three interlinked levels to cope with a wide range of error types present in the input. Each level corrects different types of errors; links between the levels allow capturing errors in word order and complex discontinuous expressions. Errors are not only corrected, but also classified. The annotation scheme is tested on a doubly-annotated sample of approx. 10,000 words with fair inter-annotator agreement results. We also explore options of application of automated linguistic annotation tools (taggers, spell checkers and grammar checkers) on the learner text to support or even substitute manual annotation.

Pedagogical stances and their multimodal signals.

Isabella Poggi, Francesca D'Errico and Giovanna Leone

The paper defines the notion of "pedagogical stance", viewed as the type of position taken, the role assumed, the image projected and the types of social behaviours performed by a teacher in her teaching interaction with a pupil. Two aspects of pedagogical stance, "didactic" and "affective – relational", are distinguished and a hypothesis is put forward about their determinant factors (the teacher's personality, idea of one's role and of the learning process, and model of the pupil). Based on a qualitative analysis of the verbal and bodily behaviour of teachers in a corpus of teacher-pupil interactions, the paper singles out two didactic stances (maieutic and efficient) and four affective-relational ones (friendly, dominating, paternalistic, and secure base). Some examples of these stances are analysed in detail and the respective patterns of verbal and behavioural signals that typically characterize the six types of stances are outlined.

Annotated Corpora for Word Alignment between Japanese and English and its Evaluation with MAP-based Word Aligner

Tsuyoshi Okita

This paper presents two annotated corpora for word alignment between Japanese and English. We annotated on top of the IWSLT-2006 and the NTCIR-8 corpora. The IWSLT-2006 corpus is in the domain of travel conversation while the NTCIR-8 corpus is in the domain of patent. We annotated the first 500 sentence pairs from the IWSLT-2006 corpus and the first 100 sentence

pairs from the NTCIR-8 corpus. After mentioned the annotation guideline, we present two evaluation algorithms how to use such hand-annotated corpora: although one is a well-known algorithm for word alignment researchers, one is novel which intends to evaluate a MAP-based word aligner of Okita et al. (2010b).

Ubiquitous Usage of a Broad Coverage French Corpus: Processing the Est Republicain corpus

Djamé Seddah, Marie Candito, Benoit Crabbé and Enrique Henestroza Anguiano

In this paper, we introduce a set of resources that we have derived from the EST RÉPUBLICAIN CORPUS, a large, freely-available collection of regional newspaper articles in French, totaling 150 million words. Our resources are the result of a full NLP treatment of the EST RÉPUBLICAIN CORPUS: handling of multi-word expressions, lemmatization, part-of-speech tagging, and syntactic parsing. Processing of the corpus is carried out using statistical machine-learning approaches - joint model of data driven lemmatization and part- of-speech tagging, PCFG-LA and dependency based models for parsing - that have been shown to achieve state-of-the-art performance when evaluated on the French Treebank. Our derived resources are made freely available, and released according to the original Creative Common license for the EST RÉPUBLICAIN CORPUS. We additionally provide an overview of the use of these resources in various applications, in particular the use of generated word clusters from the corpus to alleviate lexical data sparseness for statistical parsing.

P35 - Language Resource Infrastructures (2)

Friday, May 25, 9:45

Chairperson: **Claudia Soria** Poster Session

Federated Search: Towards a Common Search Infrastructure

Herman Stehouwer, Matej Durco, Eric Auer and Daan Broeder

Within scientific institutes there exist many language resources. These resources are often quite specialized and relatively unknown. The current infrastructural initiatives try to tackle this issue by collecting metadata about the resources and establishing centers with stable repositories to ensure the availability of the resources. It would be beneficial if the researcher could, by means of a simple query, determine which resources and which centers contain information useful to his or her research, or even work on a set of distributed resources as a virtual corpus. In this article we propose an architecture for a distributed search environment allowing researchers to perform searches in a set of distributed language resources.

Proper Language Resource Centers

Willem Elbers, Daan Broeder and Dieter van Uytvanck

Language resource centers allow researchers to reliably deposit their structured data together with associated meta data and run services operating on this deposited data. We are looking into possibilities to create long-term persistency of both the deposited data and the services operating on this data. Challenges, both technical and non-technical, that need to be solved are the need to replicate more than just the data, proper identification of the digital objects in a distributed environment by making use of persistent identifiers and the set-up of a proper authentication and authorization domain including the management of the authorization information on the digital objects. We acknowledge the investment that most language resource centers have made in their current infrastructure. Therefore one of the most important requirements is the loose coupling with existing infrastructures without the need to make many changes. This shift from a single language resource center into a federated environment of many language resource centers is discussed in the context of a real world center: The Language Archive supported by the Max Planck Institute for Psycholinguistics.

The Language Archive – a new hub for language resources

Sebastian Drude, Daan Broeder, Paul Trilsbeek and Peter Wittenburg

This contribution presents “The Language Archive” (TLA), a new unit at the MPI for Psycholinguistics, discussing the current developments in management of scientific data, considering the need for new data research infrastructures. Although several initiatives worldwide in the realm of language resources aim at the integration, preservation and mobilization of research data, the state of such scientific data is still often problematic. Data are often not well organized and archived and not described by metadata – even unique data such as field-work observational data on endangered languages is still mostly on perishable carriers. New data centres are needed that provide trusted, quality-reviewed, persistent services and suitable tools and that take legal and ethical issues seriously. The CLARIN initiative has established criteria for suitable centres. TLA is in a good position to be one of such centres. It is based on three essential pillars: (1) A data archive; (2) management, access and annotation tools; (3) archiving and software expertise for collaborative projects. The archive hosts mostly observational data on small languages

worldwide and language acquisition data, but also data resulting from experiments.

LAMP: A Multimodal Web Platform for Collaborative Linguistic Analysis

Kais Dukes and Eric Atwell

This paper describes the underlying software platform used to develop and publish annotations for the Quranic Arabic Corpus (QAC). The QAC (Dukes, Atwell and Habash, 2011) is a multimodal language resource that integrates deep tagging, interlinear translation, multiple speech recordings, visualization and collaborative analysis for the Classical Arabic language of the Quran. Available online at http://corpus.quran.com, the website is a popular study guide for Quranic Arabic, used by over 1.2 million visitors over the past year. We provide a description of the underlying software system that has been used to develop the corpus annotations. The multimodal data is made available online through an accessible cross-referenced web interface. Although our Linguistic Analysis Multimodal Platform (LAMP) has been applied to the Classical Arabic language of the Quran, we argue that our annotation model and software architecture may be of interest to other related corpus linguistics projects. Work related to LAMP includes recent efforts for annotating other Classical languages, such as Ancient Greek and Latin (Bamman, Mambrini and Crane, 2009), as well as commercial systems (e.g. Logos Bible study) that provide access to syntactic tagging for the Hebrew Bible and Greek New Testament (Brannan, 2011).

An NLP Curator (or: How I Learned to Stop Worrying and Love NLP Pipelines)

James Clarke, Vivek Srikumar, Mark Sammons and Dan Roth

Natural Language Processing continues to grow in popularity in a range of research and commercial applications, yet managing the wide array of potential NLP components remains a difficult problem. This paper describes Curator, an NLP management framework designed to address some common problems and inefficiencies associated with building NLP process pipelines; and Edison, an NLP data structure library in Java that provides streamlined interactions with Curator and offers a range of useful supporting functionality.

Using Language Resources in Humanities research

Marta Villegas, Núria Bel, Carlos Gonzalo, Amparo Moreno and Nuria Simelio

In this paper we present two real cases, in the fields of discourse analysis of newspapers and communication research which demonstrate the impact of Language Resources (LR) and NLP in the humanities. We describe our collaboration with (i) the Feminario research group from the UAB which has been investigating androcentric practices in Spanish general press since the 80s and whose research suggests that Spanish general press has undergone a dehumanization process that excludes women and men and (ii) the "Municipals'11 online" project which investigates the Spanish local election campaign in the blogosphere. We will see how NLP tools and LRs make possible the so called 'e-Humanities research' as they provide Humanities with tools to perform intensive and automatic text analyses. Language technologies have evolved a lot and are mature enough to provide useful tools to researchers dealing with large amount of textual data. The language resources that have been developed within the field of NLP have proven to be useful for other disciplines that are unaware of their existence and nevertheless would greatly benefit from them as they provide (i) exhaustiveness -to guarantee that data coverage is wide and representative enough- and (ii) reliable and significant results -to guarantee that the reported results are statistically significant.

Glottolog/Langdoc:Increasing the visibility of grey literature for low-density languages

Sebastian Nordhoff and Harald Hammarström

Language resources can be divided into structural resources treating phonology, morphosyntax, semantics etc. and resources treating the social, demographic, ethnic, political context. A third type are meta-resources, like bibliographies, which provide access to the resources of the first two kinds. This poster will present the Glottolog/Langdoc project, a comprehensive bibliography providing web access to 180k bibliographical records to (mainly) low visibility resources from low-density languages. The resources are annotated for macro-area, content language, and document type and are available in XHTML and RDF.

The Australian National Corpus: National Infrastructure for Language Resources

Steve Cassidy, Michael Haugh, Pam Peters and Mark Fallu

The Australian National Corpus has been established in an effort to make currently scattered and relatively inaccessible data available to researchers through an online portal. In contrast to other national corpora, it is conceptualised as a linked collection of many existing and future language resources representing language use in Australia, unified through common technical standards. This approach allows us to bootstrap a significant collection and add value to existing resources by providing a unified, online tool-set to support research in a number of

disciplines. This paper provides an outline of the technical platform being developed to support the corpus and a brief overview of some of the collections that form part of the initial version of the Australian National Corpus.

META-SHARE v2: An Open Network of Repositories for Language Resources including Data and Tools

Christian Federmann, Ioanna Giannopoulou, Christian Girardi, Olivier Hamon, Dimitris Mavroeidis, Salvatore Minutoli and Marc Schröder

We describe META-SHARE which aims at providing an open, distributed, secure, and interoperable infrastructure for the exchange of language resources, including both data and tools. The application has been designed and is developed as part of the T4ME Network of Excellence. We explain the underlying motivation for such a distributed repository for metadata storage and give a detailed overview on the META-SHARE application and its various components. This includes a discussion of the technical architecture of the system as well as a description of the component-based metadata schema format which has been developed in parallel. Development of the META-SHARE infrastructure adopts state-of-the-art technology and follows an open-source approach, allowing the general community to participate in the development process. The META-SHARE software package including full source code has been released to the public in March 2012. We look forward to present an up-to-date version of the META-SHARE software at the conference.

Linguagrid: a network of Linguistic and Semantic Services for the Italian Language.

Alessio Bosca, Luca Dini, Milen Kouylekov and Marco Trevisan

In order to handle the increasing amount of textual information today available on the web and exploit the knowledge latent in this mass of unstructured data, a wide variety of linguistic knowledge and resources (Language Identification, Morphological Analysis, Entity Extraction, etc.). is crucial. In the last decade LRaas (Language Resource as a Service) emerged as a novel paradigm for publishing and sharing these heterogeneous software resources over the Web. In this paper we present an overview of Linguagrid, a recent initiative that implements an open network of linguistic and semantic Web Services for the Italian language, as well as a new approach for enabling customizable corpus-based linguistic services on Linguagrid LRaaS infrastructure. A corpus ingestion service in fact allows users to upload corpora of documents and to generate classification/clustering models tailored to their needs by means of standard machine learning techniques applied to the textual contents and metadata from the corpora. The models so generated can then be accessed through proper Web Services and exploited to process and classify new textual contents.

P36 - Speech Synthesis

Friday, May 25, 9:45

Chairperson: **Martine Garnier-Rizet** Poster Session

Versatile Speech Databases for High Quality Synthesis for Basque

Iñaki Sainz, Daniel Erro, Eva Navas, Inma Hernáez, Jon Sánchez, Ibon Saratxaga and Igor Odriozola

This paper presents three new speech databases for standard Basque. They are designed primarily for corpus-based synthesis but each database has its specific purpose: 1) AhoSyn: high quality speech synthesis (recorded also in Spanish), 2) AhoSpeakers: voice conversion and 3) AhoEmo3: emotional speech synthesis. The whole corpus design and the recording process are described with detail. Once the databases were collected all the data was automatically labelled and annotated. Then, an HMM-based TTS voice was built and subjectively evaluated. The results of the evaluation are pretty satisfactory: 3.70 MOS for Basque and 3.44 for Spanish. Therefore, the evaluation assesses the quality of this new speech resource and the validity of the automated processing presented.

Building a synchronous corpus of acoustic and 3D facial marker data for adaptive audio-visual speech synthesis

Dietmar Schabus, Michael Pucher and Gregor Hofer

We have created a synchronous corpus of acoustic and 3D facial marker data from multiple speakers for adaptive audio-visual text-to-speech synthesis. The corpus contains data from one female and two male speakers and amounts to 223 Austrian German sentences each. In this paper, we first describe the recording process, using professional audio equipment and a marker-based 3D facial motion capturing system for the audio-visual recordings. We then turn to post-processing, which incorporates forced alignment, principal component analysis (PCA) on the visual data, and some manual checking and corrections. Finally, we describe the resulting corpus, which will be released under a research license at the end of our project. We show that the standard PCA based feature extraction approach also works on a multi-speaker database in the adaptation scenario, where there is no data from the target speaker available in the PCA step.

Building Text-To-Speech Voices in the Cloud

Alistair Conkie, Thomas Okken, Yeon-Jun Kim and Giuseppe Di Fabbrizio

The AT&T VoiceBuilder provides a new tool to researchers and practitioners who want to have their voices synthesized by a high-quality commercial-grade text-to-speech system without the need to install, configure, or manage speech processing software and equipment. It is implemented as a web service on the AT&T Speech Mashup Portal.The system records and validates users' utterances, processes them to build a synthetic voice and provides a web service API to make the voice available to real-time applications through a scalable cloud-based processing platform. All the procedures are automated to avoid human intervention. We present experimental comparisons of voices built using the system.

Building Synthetic Voices in the META-NET Framework

Emília Garcia Casademont, Antonio Bonafonte and Asunción Moreno

METANET4U is a European project aiming at supporting language technology for European languages and multilingualism. It is a project in the META-NET Network of Excellence, a cluster of projects aiming at fostering the mission of META, which is the Multilingual Europe Technology Alliance, dedicated to building the technological foundations of a multilingual European information society. This paper describe the resources produced at our lab to provide Synthethic voices. Using existing 10h corpus for a male and a female Spanish speakers, voices have been developed to be used in Festival, both with unit-selection and with statistical-based technologies. Furthermore, using data produced for supporting research on intra and inter-lingual voice conversion, four bilingual voices (English/Spanish) have been developed. The paper describes these resources which are available through META. Furthermore, an evaluation is presented to compare different synthesis techniques, influence of amount of data in statistical speech synthesis and the effect of sharing data in bilingual voices.

Building Text-to-Speech Systems for Resource Poor Languages

Nur-Hana Samsudin and Mark Lee

This paper describes research on building text-to-speech synthesis systems (TTS) for resource poor languages using available resources from other languages and describes our general approach to building cross-linguistic polyglot TTS. Our approach involves three main steps: language clustering, grapheme to phoneme mapping and prosody modelling. We have tested the mapping of phonemes from German to English and from Indonesian to Spanish. We have also constructed three prosody representations for different language characteristics. For evaluation we have developed an English TTS based on German data, and a Spanish TTS based on Indonesian data and compared their performance against pre-existing monolingual TTSs. Since our motivation is to develop speech synthesis for resource poor languages, we have also developed three TTS for Iban, an Austronesian language with practically no available language resources, using Malay, Indonesian and Spanish resources.

Evaluating expressive speech synthesis from audiobook corpora for conversational phrases

Eva Szekely, Joao Paulo Cabral, Mohamed Abou-Zleikha, Peter Cahill and Julie Carson-Berndsen

Audiobooks are a rich resource of large quantities of natural sounding, highly expressive speech. In our previous research we have shown that it is possible to detect different expressive voice styles represented in a particular audiobook, using unsupervised clustering to group the speech corpus of the audiobook into smaller subsets representing the detected voice styles. These subsets of corpora of different voice styles reflect the various ways a speaker uses their voice to express involvement and affect, or imitate characters. This study is an evaluation of the detection of voice styles in an audiobook in the application of expressive speech synthesis. A further aim of this study is to investigate the usability of audiobooks as a language resource for expressive speech synthesis of utterances of conversational speech. Two evaluations have been carried out to assess the effect of the genre transfer: transmitting expressive speech from read aloud literature to conversational phrases with the application of speech synthesis. The first evaluation revealed that listeners have different voice style preferences for a particular conversational phrase. The second evaluation showed that it is possible for users of speech synthesis systems to learn the characteristics of a voice style well enough to make reliable predictions about what a certain utterance will sound like when synthesised using that voice style.

P37 - Speech Resources

Friday, May 25, 9:45

Chairperson: **Andrea Paoloni** Poster Session

Body-conductive acoustic sensors in human-robot communication

Panikos Heracleous, Carlos Ishi, Takahiro Miyashita and Norihiro Hagita

In this study, the use of alternative acoustic sensors in human-robot communication is investigated. In particular, a Non-

Audible Murmur (NAM) microphone was applied in teleoperating Geminoid HI-1 robot in noisy environments. The current study introduces the methodology and the results of speech intelligibility subjective tests when a NAM microphone was used in comparison with using a standard microphone. The results show the advantage of using NAM microphone when the operation takes place in adverse environmental conditions. In addition, the effect of Geminoid's lip movements on speech intelligibility is also investigated. Subjective speech intelligibility tests show that the operator's speech can be perceived with higher intelligibility scores when operator's audio speech is perceived along with the lip movements of robots.

Balanced data repository of spontaneous spoken Czech

Lucie Válková, Martina Waclawičová and Michal Křen

The paper presents data repository that will be used as a source of data for ORAL2013, a new corpus of spontaneous spoken Czech. The corpus is planned to be published in 2013 within the framework of the Czech National Corpus and it will contain both the audio recordings and their transcriptions manually aligned with time stamps. The corpus will be designed as a representation of contemporary spontaneous spoken language used in informal, real-life situations on the area of the whole Czech Republic and thus balanced in the main sociolinguistic categories of speakers. Therefore, the data repository features broad regional coverage with large variety of speakers, as well as precise and uniform processing. The repository is already built, basically balanced and sized 3 million words proper (i.e. tokens not including punctuation). Before the publication, another set of overall consistency checks will be carried out, as well as final selection of the transcriptions to be included into ORAL2013 as the final product.

NKI-CCRT Corpus - Speech Intelligibility Before and After Advanced Head and Neck Cancer Treated with Concomitant Chemoradiotherapy

R.P. Clapham, L. van der Molen, R.J.JH. van Son, M. van den Brekel and F.J.M. Hilgers

Evaluations of speech intelligibility based on a read passage are often used in the clinical situation to assess the impact of the disease and/or treatment on spoken communication. Although scale-based measures are often used in the clinical setting, these measures are susceptible to listener response bias. Automatic evaluation tools are being developed in response to some of the drawbacks of perceptual evaluation, however, large corpora judged by listeners are needed to improve and test these tools. To this end, the NKI-CCRT corpus with individual listener judgements on the intelligibility of recordings of 55 speakers treated for cancer of the head and neck will be made available for restricted scientific use. The corpus contains recordings and perceptual evaluations of speech intelligibility over three evaluation moments: before treatment and after treatment (10-weeks and 12-months). Treatment was by means of chemoradiotherapy (CCRT). Thirteen recently graduated speech pathologists rated the speech intelligibility of the recordings on a 7-point scale. Information on recording and perceptual evaluation procedures is presented in addition to preliminary rater reliability and agreement information. Preliminary results show that for many speakers speech intelligibility is rated low before cancer treatment.

Sense Meets Nonsense - Sense Meets Nonsense - a dual-layer Danish speech corpus for perception studies

Thomas Ulrich Christiansen and Peter Juel Henrichsen

In this paper, we present the newly established Danish speech corpus PiTu. The corpus consists of recordings of 28 native Danish talkers (14 female and 14 male) each reproducing (i) a series of nonsense syllables, and (ii) a set of authentic natural language sentences. The speech corpus is tailored for investigating the relationship between early stages of the speech perceptual process and later stages. We present our considerations involved in preparing the experimental set-up, producing the anechoic recordings, compiling the data, and exploring the materials in linguistic research. We report on a small pilot experiment demonstrating how PiTu and similar speech corpora can be used in studies of prosody as a function of semantic content. The experiment addresses the issue of whether the governing principles of Danish prosody assignment is mainly talker-specific or mainly content-typical (under the specific experimental conditions). The corpus is available in its entirety for download at http://amtoolbox.sourceforge.net/pitu/.

SMALLWorlds – Multilingual Content-Controlled Monologues

Peter Juel Henrichsen and Marcus Uneson

We present the speech corpus SMALLWorlds (Spoken Multi-lingual Accounts of Logically Limited Worlds), newly established and still growing. SMALLWorlds contains monologic descriptions of scenes or worlds which are simple enough to be formally describable. The descriptions are instances of content-controlled monologue: semantically"pre-specified"but still bearing most hallmarks of spontaneous speech (hesitations and filled pauses, relaxed syntax, repetitions, self-corrections, incomplete constituents, irrelevant or redundant information, etc.)

as well as idiosyncratic speaker traits. In the paper, we discuss the pros and cons of data so elicited. Following that, we present a typical SMALLWorlds task: the description of a simple drawing with differently coloured circles, squares, and triangles, with no hints given as to which description strategy or language style to use. We conclude with an example on how SMALLWorlds may be used: unsupervised lexical learning from phonetic transcription. At the time of writing, SMALLWorlds consists of more than 250 recordings in a wide range of typologically diverse languages from many parts of the world, some unwritten and endangered.

A Parameterized and Annotated Spoken Dialog Corpus of the CMU Let's Go Bus Information System

Alexander Schmitt, Stefan Ultes and Wolfgang Minker

Standardized corpora are the foundation for spoken language research. In this work, we introduce an annotated and standardized corpus in the Spoken Dialog Systems (SDS) domain. Data from the Let's Go Bus Information System from the Carnegie Mellon University in Pittsburgh has been formatted, parameterized and annotated with quality, emotion, and task success labels containing 347 dialogs with 9,083 system-user exchanges. A total of 46 parameters have been derived automatically and semi-automatically from Automatic Speech Recognition (ASR), Spoken Language Understanding (SLU) and Dialog Manager (DM) properties. To each spoken user utterance an emotion label from the set garbage, non-angry, slightly angry, very angry has been assigned. In addition, a manual annotation of Interaction Quality (IQ) on the exchange level has been performed with three raters achieving a Kappa value of 0.54. The IQ score expresses the quality of the interaction up to each system-user exchange on a score from 1-5. The presented corpus is intended as a standardized basis for classification and evaluation tasks regarding task success prediction, dialog quality estimation or emotion recognition to foster comparability between different approaches on these fields.

Speech and Language Resources for LVCSR of Russian

Sergey Zablotskiy, Alexander Shvets, Maxim Sidorov, Eugene Semenkin and Wolfgang Minker

A syllable-based language model reduces the lexicon size by hundreds of times. It is especially beneficial in case of highly inflective languages like Russian due to the abundance of word forms according to various grammatical categories. However, the main arising challenge is the concatenation of recognised syllables into the originally spoken sentence or phrase, particularly in the presence of syllable recognition mistakes. Natural fluent speech does not usually incorporate clear information about the outside borders of the spoken words. In this paper a method for the syllable concatenation and error correction is suggested and tested. It is based on the designed co-evolutionary asymptotic probabilistic genetic algorithm for the determination of the most likely sentence corresponding to the recognized chain of syllables within an acceptable time frame. The advantage of this genetic algorithm modification is the minimum number of settings to be manually adjusted comparing to the standard algorithm. Data used for acoustic and language modelling are also described here. A special issue is the preprocessing of the textual data, particularly, handling of abbreviations, Arabic and Roman numerals, since their inflection mostly depends on the context and grammar.

Dysarthric Speech Database for Development of QoLT Software Technology

Dae-Lim Choi, Bong-Wan Kim, Yeon-Whoa Kim, Yong-Ju Lee, Yongnam Um and Minhwa Chung

This paper describes the creation of a dysarthric speech database which has been done as part of a national program to help the disabled lead a better life – a challenging endeavour that is targeting development of speech technologies for people with articulation disabilities. The additional aims of this database are to study the phonetic characteristics of the different types of the disabled persons, develop the automatic method to assess degrees of disability, and investigate the phonetic features of dysarthric speech. For these purposes, a large database of about 600 mildly or moderately severe dysarthric persons is planned for a total of 4 years (2010.06. 01 2014.05.31). At present a dysarthric speech database of 120 speakers has been collected and we are continuing to record new speakers with cerebral paralysis of mild and moderate severity. This paper also introduces the prompting items, the assessment of the speech disability severity of the speakers, and other considerations for the creation of a dysarthric speech.

The annotation of the C-ORAL-BRASIL oral through the implementation of the Palavras Parser

Eckhard Bick, Heliana Mello, Alessandro Panunzi and Tommaso Raso

This article describes the morphosyntactic annotation of the C-ORAL-BRASIL speech corpus, using an adapted version of the Palavras parser. In order to achieve compatibility with annotation rules designed for standard written Portuguese, transcribed words were orthographically normalized, and the parsing lexicon augmented with speech-specific material, phonetically spelled

abbreviations etc. Using a two-level annotation approach, speech flow markers like overlaps, retractions and non-verbal productions were separated from running, annotatable text. In the absence of punctuation, syntactic segmentation was achieved by exploiting prosodic break markers, enhanced by a rule-based distinctions between pause and break functions. Under optimal conditions, the modified parsing system achieved correctness rates (F-scores) of 98.6% for part of speech, 95% for syntactic function and 99% for lemmatization. Especially at the syntactic level, a clear connection between accessibility of prosodic break markers and annotation performance could be documented.

The Nordic Dialect Corpus

Janne Bondi Johannessen, Joel Priestley, Kristin Hagen, Anders Nøklestad and André Lynum

In this paper, we describe the Nordic Dialect Corpus, which has recently been completed. The corpus has a variety of features that combined makes it an advanced tool for language researchers. These features include: Linguistic contents (dialects from five closely related languages), annotation (tagging and two types of transcription), search interface (advanced possibilities for combining a large array of search criteria and results presentation in an intuitive and simple interface), many search variables (linguistics-based, informant-based, time-based), multimedia display (linking of sound and video to transcriptions), display of results in maps, display of informant details (number of words and other information on informants), advanced results handling (concordances, collocations, counts and statistics shown in a variety of graphical modes, plus further processing). Finally, and importantly, the corpus is freely available for research on the web. We give examples of both various kinds of searches, of displays of results and of results handling.

ULex: new data models and a mobile environment for corpus enrichment.

Dafydd Gibbon

The Ubiquitous Lexicon concept (ULex) has two sides. In the first kind of ubiquity, ULex combines prelexical corpus based lexicon extraction and formatting techniques from speech technology and corpus linguistics for both language documentation and basic speech technology (e.g. speech synthesis), and proposes new XML models for the basic datatypes concerned, in order to enable standardisastion and data interchange in these areas. The prelexical data types range from basic wordlists through diphone tables to concordance and interlinear glossing structures. While several proposals for standardising XML models of lexicon types are available, these more basic pre-lexical, data types, which are important in lexical acquisition, have received little attention. In the second area of ubiquity, ULex is implemented in a novel mobile environment to enable collaborative cross-platform use via a web application, either on the internet or, via a local hotspot, on an intranet, which runs not only on standard PC types but also on tablet computers and smartphones and is thereby also rendered truly ubiquitous in a geographical sense.

Developing Partially-Transcribed Speech Corpus from Edited Transcriptions

Kengo Ohta, Masatoshi Tsuchiya and Seiichi Nakagawa

Large-scale spontaneous speech corpora are crucial resource for various domains of spoken language processing. However, the available corpora are usually limited because their construction cost is quite expensive especially in transcribing speech precisely. On the other hand, loosely transcribed corpora like shorthand notes, meeting records and closed captions are more widely available than precisely transcribed ones, because their imperfectness reduces their construction cost. Because these corpora contain both precisely transcribed regions and edited regions, it is difficult to use them directly as speech corpora for learning acoustic models. Under this background, we have been considering to build an efficient semi-automatic framework to convert loose transcriptions to precise ones. This paper describes an improved automatic detection method of precise regions from loosely transcribed corpora for the above framework. Our detection method consists of two steps: the first step is a force alignment between loose transcriptions and their utterances to discover the corresponding utterance for the certain loose transcription, and the second step is a detector of precise regions with a support vector machine using several features obtained from the first step. Our experimental result shows that our method achieves a high accuracy of detecting precise regions, and shows that the precise regions extracted by our method are effective as training labels of lightly supervised speaker adaptation.

LDC Forced Aligner

Xiaoyi Ma

This paper describes the LDC forced aligner which was designed to align audio and transcripts. Unlike existing forced aligners, LDC forced aligner can align partially transcribed audio files, and also audio files with large chunks of non-speech segments, such as noise, music, silence etc, by inserting optional wildcard phoneme sequences between sentence or paragraph boundaries. Based on the HTK tool kit, LDC forced aligner can align audio and transcript on sentence or word level. This paper also reports

its usage on English and Mandarin Chinese data.

The KIT Lecture Corpus for Speech Translation

Sebastian Stüker, Florian Kraft, Christian Mohr, Teresa Herrmann, Eunah Cho and Alex Waibel

Academic lectures offer valuable content, but often do not reach their full potential audience due to the language barrier. Human translations of lectures are too expensive to be widely used. Speech translation technology can be an affordable alternative in this case. State-of-the-art speech translation systems utilize statistical models that need to be trained on large amounts of in-domain data. In order to support the KIT lecture translation project in its effort to introduce speech translation technology in KIT's lecture halls, we have collected a corpus of German lectures at KIT. In this paper we describe how we recorded the lectures and how we annotated them. We further give detailed statistics on the types of lectures in the corpus and its size. We collected the corpus with the purpose in mind that it should not just be suited for training a spoken language translation system the traditional way, but should also enable us to research techniques that enable the translation system to automatically and autonomously adapt itself to the varying topics and speakers of lectures

Development of Text and Speech database for Hindi and Indian English specific to Mobile Communication environment

Shyam Agrawal, Shweta Sinha, Pooja Singh and Jesper Olson

Abstract This paper describes the method and experiences of text and speech data collection in mobile communication in Indian English Hindi. The primary data collection is done in the form of large number of messages as part of Personal communication among natives of Hindi language and Indian speakers of English. To gather the versatility of mobile communication database among Hindi and English, 12 domains were identified for collection of text corpus from speaking population belonging to deferent age groups, sex and dialects. The text obtained in raw form based on slangs and unconventional grammar were cleaned using on language grammar rules and then tagged and expanded to explain context specific meaning of the words. Texts of 1163 participants from Hindi speaking regions and 1405 English users were taken for creating 13 prompt sheets; containing 630 phonetically rich sentences created using a special software. Each prompt sheet was recorded by at least 7 users simultaneously in three channels and recorded by a total of 100 speakers and annotated. The work is a step forward in the direction of development of standards for mobile text and speech data collection for Indian languages. Keywords - Speech data base, Text analysis, mobile communication, Hindi and Indian English Speech, multi-lingual speech processing.

O41 - Machine Translation and Language Resources (2)

Friday, May 25, 11:45

Chairperson: **Atsushi Fujii** Oral Session

Source-Language Dictionaries Help Non-Expert Users to Enlarge Target-Language Dictionaries for Machine Translation

Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis and Juan Antonio Pérez-Ortiz

In this paper, a previous work on the enlargement of monolingual dictionaries of rule-based machine translation systems by non-expert users is extended to tackle the complete task of adding both source-language and target-language words to the monolingual dictionaries and the bilingual dictionary. In the original method, users validate whether some suffix variations of the word to be inserted are correct in order to find the most appropriate inflection paradigm. This method is now improved by taking advantage from the strong correlation detected between paradigms in both languages to reduce the search space of the target-language paradigm once the source-language paradigm is known. Results show that, when the source-language word has already been inserted, the system is able to more accurately predict which is the right target-language paradigm, and the number of queries posed to users is significantly reduced. Experiments also show that, when the source language and the target language are not closely related, it is only the source-language part-of-speech category, but not the rest of information provided by the source-language paradigm, which helps to correctly classify the target-language word.

The ML4HMT Workshop on Optimising the Division of Labour in Hybrid Machine Translation

Christian Federmann, Eleftherios Avramidis, Marta R. Costa-Jussà, Josef van Genabith, Maite Melero and Pavel Pecina

We describe the "Shared Task on Applying Machine Learning Techniques to Optimise the Division of Labour in Hybrid Machine Translation" (ML4HMT) which aims to foster research on improved system combination approaches for machine translation (MT). Participants of the challenge are requested to build hybrid translations by combining the output of several MT systems of different types. We first describe the ML4HMT corpus used in the shared task, then explain the XLIFF-based annotation format

we have designed for it, and briefly summarize the participating systems. Using both automated metrics scores and extensive manual evaluation, we discuss the individual performance of the various systems. An interesting result from the shared task is the fact that we were able to observe different systems winning according to the automated metrics scores when compared to the results from the manual evaluation. We conclude by summarising the first edition of the challenge and by giving an outlook to future work.

Alignment-based reordering for SMT

Maria Holmqvist, Sara Stymne, Lars Ahrenberg and Magnus Merkel

We present a method for improving word alignment quality for phrase-based statistical machine translation by reordering the source text according to the target word order suggested by an initial word alignment. The reordered text is used to create a second word alignment which can be an improvement of the first alignment, since the word order is more similar. The method requires no other pre-processing such as part-of-speech tagging or parsing. We report improved Bleu scores for English-to-German and English-to-Swedish translation. We also examined the effect on word alignment quality and found that the reordering method increased recall while lowering precision, which partly can explain the improved Bleu scores. A manual evaluation of the translation output was also performed to understand what effect our reordering method has on the translation system. We found that where the system employing reordering differed from the baseline in terms of having more words, or a different word order, this generally led to an improvement in translation quality.

Same domain different discourse style - A case study on Language Resources for data-driven Machine Translation

Monica Gavrila, Walther v. Hahn and Cristina Vertan

Data-driven machine translation (MT) approaches became very popular during last years, especially for language pairs for which it is difficult to find specialists to develop transfer rules. Statistical (SMT) or example-based (EBMT) systems can provide reasonable translation quality for assimilation purposes, as long as a large amount of training data is available. Especially SMT systems rely on parallel aligned corpora which have to be statistical relevant for the given language pair. The construction of large domain specific parallel corpora is time- and cost-consuming; the current practice relies on one or two big such corpora per language pair. Recent developed strategies ensure certain portability to other domains through specialized lexicons or small domain specific corpora. In this paper we discuss the influence of different discourse styles on statistical machine translation systems. We investigate how a pure SMT performs when training and test data belong to same domain but the discourse style varies.

Automatic Translation of Scholarly Terms into Patent Terms Using Synonym Extraction Techniques

Hidetsugu Nanba, Toshiyuki Takezawa, Kiyoko Uchiyama and Akiko Aizawa

Retrieving research papers and patents is important for any researcher assessing the scope of a field with high industrial relevance. However, the terms used in patents are often more abstract or creative than those used in research papers, because they are intended to widen the scope of claims. Therefore, a method is required for translating scholarly terms into patent terms. In this paper, we propose six methods for translating scholarly terms into patent terms using two synonym extraction methods: a statistical machine translation (SMT)-based method and a distributional similarity (DS)-based method. We conducted experiments to confirm the effectiveness of our method using the dataset of the Patent Mining Task from the NTCIR-7 Workshop. The aim of the task was to classify Japanese language research papers (pairs of titles and abstracts) using the IPC system at the subclass (third level), main group (fourth level), and subgroup (the fifth and most detailed level). The results showed that an SMT-based method (SMT_ ABST+IDF) performed best at the subgroup level, whereas a DS-based method (DS+IDF) performed best at the subclass level.

O42 - WordNets

Friday, May 25, 11:45

Chairperson: **Martha Palmer** Oral Session

Towards a richer wordnet representation of properties

Sanni Nimb and Bolette Sandford Pedersen

This paper discusses how information on properties in a currently developed Danish thesaurus can be transferred to the Danish wordnet, DanNet, and in this way enrich the wordnet with the highly relevant links between properties and their external arguments (i.e. tasty – food). In spite of the fact that the thesaurus is still under development (two thirds still to be compiled) we perform an automatic transfer of relations from the thesaurus to the wordnet which shows promising results. In all, 2,362 property relations are automatically transferred to DanNet and 2% of the transferred material is manually validated. The pilot validation indicates that approx. 90 % of the transferred relations are correctly assigned whereas around 10% are either erroneous

or just not very informative, a fact which, however, can partly be explained by the incompleteness of the material at its current stage. As a further consequence, the experiment has led to a richer specification of the editor guidelines to be used in the last compilation phase of the thesaurus.

A proposal for improving WordNet Domains

Aitor Gonzalez-Agirre, Mauro Castillo and German Rigau

WordNet Domains (WND) is a lexical resource where synsets have been semi-automatically annotated with one or more domain labels from a set of 165 hierarchically organized domains. The uses of WND include the power to reduce the polysemy degree of the words, grouping those senses that belong to the same domain. But the semi-automatic method used to develop this resource was far from being perfect. By cross-checking the content of the Multilingual Central Repository (MCR) it is possible to find some errors and inconsistencies. Many are very subtle. Others, however, leave no doubt. Moreover, it is very difficult to quantify the number of errors in the original version of WND. This paper presents a novel semi-automatic method to propagate domain information through the MCR. We also compare both labellings (the original and the new one) allowing us to detect anomalies in the original WND labels.

Corpus+WordNet thesaurus generation for ontology enriching

Fernando Castilho, Roger Granada, Breno Meneghetti, Leonardo Carvalho and Renata Vieira

This paper presents a model to enrich an ontology with a thesaurus based on a domain corpus and WordNet. The model is applied to the data privacy domain and the initial domain resources comprise a data privacy ontology, a corpus of privacy laws, regulations and guidelines for projects. Based on these resources, a thesaurus is automatically generated. The thesaurus seeds are composed by the ontology concepts. For these seeds similar terms are extracted from the corpus using known thesaurus generation methods. A filtering process searches for semantic relations between seeds and similar terms within WordNet. As a result, these semantic relations are used to expand the ontology with relations between them and related terms in the corpus. The resulting resource is a hierarchical structure that can help on the ontology investigation and maintenance. The results allow the investigation of the domain knowledge with the support of semantic relations not present on the original ontology.

Cleaning noisy wordnets

Benoît Sagot and Darja Fišer

Automatic approaches to creating and extending wordnets, which have become very popular in the past decade, inadvertently result in noisy synsets. This is why we propose an approach to detect synset outliers in order to eliminate the noise and improve accuracy of the developed wordnets, so that they become more useful lexico-semantic resources for natural language applications. The approach compares the words that appear in the synset and its surroundings with the contexts of the literals in question they are used in based on large monolingual corpora. By fine-tuning the outlier threshold we can influence how many outlier candidates will be eliminated. Although the proposed approach is language-independent we test it on Slovene and French that were created automatically from bilingual resources and contain plenty of disambiguation errors. Manual evaluation of the results shows that by applying a threshold similar to the estimated error rate in the respective wordnets, 67% of the proposed outlier candidates are indeed incorrect for French and a 64% for Slovene. This is a big improvement compared to the estimated overall error rates in the resources, which are 12% for French and 15% for Slovene.

Wordnet extension made simple: A multilingual lexicon-based approach using wiki resources

Valérie Hanoka and Benoît Sagot

In this paper, we propose a simple methodology for building or extending wordnets using easily extractible lexical knowledge from Wiktionary and Wikipedia. This method relies on a large multilingual translation/synonym graph in many languages as well as synset-aligned wordnets. It guesses frequent and polysemous literals that are difficult to find using other methods by looking at back-translations in the graph, showing that the use of a heavily multilingual lexicon can be a way to mitigate the lack of wide coverage bilingual lexicon for wordnet creation or extension. We evaluate our approach on French by applying it for extending WOLF, a freely available French wordnet.

O43 - Text Mining

Friday, May 25, 11:45

Chairperson: **Paul Rayson** Oral Session

A Survey of Text Mining Architectures and the UIMA Standard

Mathias Bank and Martin Schierle

With the rising amount of digitally available text, the need for efficient processing algorithms is growing fast. Although a lot of libraries are commonly available, their modularity and interchangeability is very limited, therefore forcing a lot of reimplementations and modifications not only in research areas but also in real world application scenarios. In

recent years, different NLP frameworks have been proposed to provide an efficient, robust and convenient architecture for information processing tasks. This paper will present an overview over the most common approaches with their advantages and shortcomings, and will discuss them with respect to the first standardized architecture - the Unstructured Information Management Architecture (UIMA).

Large Scale Semantic Annotation, Indexing and Search at The National Archives

Diana Maynard and Mark A. Greenwood

This paper describes a tool developed to improve access to the enormous volume of data housed at the UK's National Archives, both for the general public and for specialist researchers. The system we have developed, TNA-Search, enables a multi-paradigm search over the entire electronic archive (42TB of data in various formats). The search functionality allows queries that arbitrarily mix any combination of full-text, structural, linguistic and semantic queries. The archive is annotated and indexed with respect to a massive semantic knowledge base containing data from the LOD cloud, data.gov.uk, related TNA projects, and a large geographical database. The semantic annotation component achieves approximately 83% F-measure, which is very reasonable considering the wide range of entities and document types and the open domain. The technologies are being adopted by real users at The National Archives and will form the core of their suite of search tools, with additional in-house interfaces.

Expertise Mining for Enterprise Content Management

Georgeta Bordea, Sabrina Kirrane, Paul Buitelaar and Bianca Pereira

Enterprise content analysis and platform configuration for enterprise content management is often carried out by external consultants that are not necessarily domain experts. In this paper, we propose a set of methods for automatic content analysis that allow users to gain a high level view of the enterprise content. Here, a main concern is the automatic identification of key stakeholders that should ideally be involved in analysis interviews. The proposed approach employs recent advances in term extraction, semantic term grounding, expert profiling and expert finding in an enterprise content management setting. Extracted terms are evaluated using human judges, while term grounding is evaluated using a manually created gold standard for the DBpedia datasource.

SemSim: Resources for Normalized Semantic Similarity Computation Using Lexical Networks

Elias Iosif and Alexandros Potamianos

We investigate the creation of corpora from web-harvested data following a scalable approach that has linear query complexity. Individual web queries are posed for a lexicon that includes thousands of nouns and the retrieved data are aggregated. A lexical network is constructed, in which the lexicon nouns are linked according to their context-based similarity. We introduce the notion of semantic neighborhoods, which are exploited for the computation of semantic similarity. Two types of normalization are proposed and evaluated on the semantic tasks of: (i) similarity judgement, and (ii) noun categorization and taxonomy creation. The created corpus along with a set of tools and noun similarities are made publicly available.

Identification of Manner in Bio-Events

Raheel Nawaz, Paul Thompson and Sophia Ananiadou

Due to the rapid growth in the volume of biomedical literature, there is an increasing requirement for high-performance semantic search systems, which allow biologists to perform precise searches for events of interest. Such systems are usually trained on corpora of documents that contain manually annotated events. Until recently, these corpora, and hence the event extraction systems trained on them, focussed almost exclusively on the identification and classification of event arguments, without taking into account how the textual context of the events could affect their interpretation. Previously, we designed an annotation scheme to enrich events with several aspects (or dimensions) of interpretation, which we term meta-knowledge, and applied this scheme to the entire GENIA corpus. In this paper, we report on our experiments to automate the assignment of one of these meta-knowledge dimensions, i.e. Manner, to recognised events. Manner is concerned with the rate, strength intensity or level of the event. We distinguish three different values of manner, i.e., High, Low and Neutral. To our knowledge, our work represents the first attempt to classify the manner of events. Using a combination of lexical, syntactic and semantic features, our system achieves an overall accuracy of 99.4%.

O44 - Evaluation of Systems and Application

Friday, May 25, 11:45

Chairperson: **Sabine Schulte im Walde** Oral Session

Cross-lingual studies of ASR errors: paradigms for perceptual evaluations

Ioana Vasilescu, Martine Adda-Decker and Lori Lamel

It is well-known that human listeners significantly outperform machines when it comes to transcribing speech. This paper presents a progress report of the joint research in the automatic vs human speech transcription and of the perceptual experiments developed at LIMSI that aims to increase our understanding

of automatic speech recognition errors. Two paradigms are described here in which human listeners are asked to transcribe speech segments containing words that are frequently misrecognized by the system. In particular, we sought to gain information about the impact of increased context to help humans disambiguate problematic lexical items, typically homophone or near-homophone words. The long-term aim of this research is to improve the modeling of ambiguous contexts so as to reduce automatic transcription errors.

Practical Evaluation of Human and Synthesized Speech for Virtual Human Dialogue Systems

Kallirroi Georgila, Alan Black, Kenji Sagae and David Traum

The current practice in virtual human dialogue systems is to use professional human recordings or limited-domain speech synthesis. Both approaches lead to good performance but at a high cost. To determine the best trade-off between performance and cost, we perform a systematic evaluation of human and synthesized voices with regard to naturalness, conversational aspect, and likability. We vary the type (in-domain vs. out-of-domain), length, and content of utterances, and take into account the age and native language of raters as well as their familiarity with speech synthesis. We present detailed results from two studies, a pilot one and one run on Amazon's Mechanical Turk. Our results suggest that a professional human voice can supersede both an amateur human voice and synthesized voices. Also, a high-quality general-purpose voice or a good limited-domain voice can perform better than amateur human recordings. We do not find any significant differences between the performance of a high-quality general-purpose voice and a limited-domain voice, both trained with speech recorded by actors. As expected, the high-quality general-purpose voice is rated higher than the limited-domain voice for out-of-domain sentences and lower for in-domain sentences. There is also a trend for long or negative-content utterances to receive lower ratings.

Designing an Evaluation Framework for Spoken Term Detection and Spoken Document Retrieval at the NTCIR-9 SpokenDoc Task

Tomoyosi Akiba, Hiromitsu Nishizaki, Kiyoaki Aikawa, Tatsuya Kawahara and Tomoko Matsui

We describe the evaluation framework for spoken document retrieval for the IR for the Spoken Documents Task, conducted in the ninth NTCIR Workshop. The two parts of this task were a spoken term detection (STD) subtask and an ad hoc spoken document retrieval subtask (SDR). Both subtasks target search terms, passages and documents included in academic and simulated lectures of the Corpus of Spontaneous Japanese. Seven teams participated in the STD subtask and five in the SDR subtask. The results obtained through the evaluation in the workshop are discussed.

Evaluation of the KomParse Conversational Non-Player Characters in a Commercial Virtual World

Tina Kluewer, Feiyu Xu, Peter Adophs and Hans Uszkoreit

The paper describes the evaluation of the KomParse system. KomParse is a dialogue system embedded in a 3-D massive multiplayer online game, allowing conversations between non player characters (NPCs) and game users. In a field test with game users, the system was evaluated with respect to acceptability and usability of the overall system as well as task completion, dialogue control and efficiency of three conversational tasks. Furthermore, subjective feedback has been collected for evaluating the single communication components of the system such as natural language understanding. The results are very satisfying and promising. In general, both the usability and acceptability tests show that the tested NPC is useful and well-accepted by the users. Even if the NPC does not always understand the users well and expresses things unexpected, he could still provide appropriate responses to help users to solve their problems or entertain them.

The IWSLT 2011 Evaluation Campaign on Automatic Talk Translation

Marcello Federico, Sebastian Stüker, Luisa Bentivogli, Michael Paul, Mauro Cettolo, Teresa Herrmann, Jan Niehues and Giovanni Moretti

We report here on the eighth evaluation campaign organized in 2011 by the IWSLT workshop series. That IWSLT 2011 evaluation focused on the automatic translation of public talks and included tracks for speech recognition, speech translation, text translation, and system combination. Unlike in previous years, all data supplied for the evaluation has been publicly released on the workshop website, and is at the disposal of researchers interested in working on our benchmarks and in comparing their results with those published at the workshop. This paper provides an overview of the IWSLT 2011 evaluation campaign, and describes the data supplied, the evaluation infrastructure made available to participants, and the subjective evaluation carried out.

P38 - Subjectivity: Sentiments, Emotions, Opinions (2)

Friday, May 25, 11:45

Chairperson: **Paolo Rosso** Poster Session

MLSA – A Multi-layered Reference Corpus for German Sentiment Analysis

Simon Clematide, Stefan Gindl, Manfred Klenner, Stefanos Petrakis, Robert Remus, Josef Ruppenhofer, Ulli Waltinger and Michael Wiegand

In this paper, we describe MLSA, a publicly available multi-layered reference corpus for German-language sentiment analysis. The construction of the corpus is based on the manual annotation of 270 German-language sentences considering three different layers of granularity. The sentence-layer annotation, as the most coarse-grained annotation, focuses on aspects of objectivity, subjectivity and the overall polarity of the respective sentences. Layer 2 is concerned with polarity on the word- and phrase-level, annotating both subjective and factual language. The annotations on Layer 3 focus on the expression-level, denoting frames of private states such as objective and direct speech events. These three layers and their respective annotations are intended to be fully independent of each other. At the same time, exploring for and discovering interactions that may exist between different layers should also be possible. The reliability of the respective annotations was assessed using the average pairwise agreement and Fleiss' multi-rater measures. We believe that MLSA is a beneficial resource for sentiment analysis research, algorithms and applications that focus on the German language.

A Classification of Adjectives for Polarity Lexicons Enhancement

Silvia Vázquez and Núria Bel

Subjective language detection is one of the most important challenges in Sentiment Analysis. Because of the weight and frequency in opinionated texts, adjectives are considered a key piece in the opinion extraction process. These subjective units are more and more frequently collected in polarity lexicons in which they appear annotated with their prior polarity. However, at the moment, any polarity lexicon takes into account prior polarity variations across domains. This paper proves that a majority of adjectives change their prior polarity value depending on the domain. We propose a distinction between domain dependent and domain independent adjectives. Moreover, our analysis led us to propose a further classification related to subjectivity degree: constant, mixed and highly subjective adjectives. Following this classification, polarity values will be a better support for Sentiment Analysis.

SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis

Jorge Carrillo de Albornoz, Laura Plaza and Pablo Gervás

This paper presents SentiSense, a concept-based affective lexicon. It is intended to be used in sentiment analysis-related tasks, specially in polarity and intensity classification and emotion identification. SentiSense attaches emotional meanings to concepts from the WordNet lexical database, instead of terms, thus allowing to address the word ambiguity problem using one of the many WordNet-based word sense disambiguation algorithms. SentiSense consists of 5,496 words and 2,190 synsets labeled with an emotion from a set of 14 emotional categories, which are related by an antonym relationship. SentiSense has been developed semi-automatically using several semantic relations between synsets in WordNet. SentiSense is endowed with a set of tools that allow users to visualize the lexicon and some statistics about the distribution of synsets and emotions in SentiSense, as well as to easily expand the lexicon. SentiSense is available for research purposes.

"Vreselijk mooi!" (terribly beautiful): A Subjectivity Lexicon for Dutch Adjectives.

Tom De Smedt and Walter Daelemans

We present a new open source subjectivity lexicon for Dutch adjectives. The lexicon is a dictionary of 1,100 adjectives that occur frequently in online product reviews, manually annotated with polarity strength, subjectivity and intensity, for each word sense. We discuss two machine learning methods (using distributional extraction and synset relations) to automatically expand the lexicon to 5,500 words. We evaluate the lexicon by comparing it to the user-given star rating of online product reviews. We show promising results in both in-domain and cross-domain evaluation. The lexicon is publicly available as part of the PATTERN software package (http://www.clips.ua.ac.be/pages/pattern).

Visualizing Sentiment Analysis on a User Forum

Rasmus Sundberg, Anders Eriksson, Johan Bini and Pierre Nugues

Sentiment analysis, or opinion mining, is the process of extracting sentiment from documents or sentences, where the expressed sentiment is typically categorized as positive, negative, or neutral. Many different techniques have been proposed. In this paper, we report the reimplementation of nine algorithms and their

evaluation across four corpora to assess the sentiment at the sentence level. We extracted the named entities from each sentence and we associated them with the sentence sentiment. We built a graphical module based on the Qlikview software suite to visualize the sentiments attached to named entities mentioned in Internet forums and follow opinion changes over time.

Affective Common Sense Knowledge Acquisition for Sentiment Analysis

Erik Cambria, Yunqing Xia and Amir Hussain

Thanks to the advent of Web 2.0, the potential for opinion sharing today is unmatched in history. Making meaning out of the huge amount of unstructured information available online, however, is extremely difficult as web-contents, despite being perfectly suitable for human consumption, still remain hardly accessible to machines. To bridge the cognitive and affective gap between word-level natural language data and the concept-level sentiments conveyed by them, affective common sense knowledge is needed. In sentic computing, the general common sense knowledge contained in ConceptNet is usually exploited to spread affective information from selected affect seeds to other concepts. In this work, besides exploiting the emotional content of the Open Mind corpus, we also collect new affective common sense knowledge through label sequential rules, crowd sourcing, and games-with-a-purpose techniques. In particular, we develop Open Mind Common Sentics, an emotion-sensitive IUI that serves both as a platform for affective common sense acquisition and as a publicly available NLP tool for extracting the cognitive and affective information associated with short texts.

P39 - Language Resource Infrastructures (3)

Friday, May 25, 11:45

Chairperson: **Penny Labropoulou** Poster Session

A Repository for the Sustainable Management of Research Data

Emanuel Dima, Verena Henrich, Erhard Hinrichs, Marie Hinrichs, Christina Hoppermann, Thorsten Trippel, Thomas Zastrow and Claus Zinn

This paper presents the system architecture as well as the underlying workflow of the Extensible Repository System of Digital Objects (ERDO) which has been developed for the sustainable archiving of language resources within the Tübingen CLARIN-D project. In contrast to other approaches focusing on archiving experts, the described workflow can be used by researchers without required knowledge in the field of long-term storage for transferring data from their local file systems into a persistent repository.

Towards a comprehensive open repository of Polish language resources

Maciej Ogrodniczuk, Piotr Pęzik and Adam Przepiórkowski

The aim of this paper is to present current efforts towards the creation of a comprehensive open repository of Polish language resources and tools (LRTs). The work described here is carried out within the CESAR project, member of the META-NET consortium. It has already resulted in the creation of the Computational Linguistics in Poland site containing an exhaustive collection of Polish LRTs. Current work is focused on the creation of new LRTs and, esp., the enhancement of existing LRTs, such as parallel corpora, annotated corpora of written and spoken Polish and morphological dictionaries to be made available via the META-SHARE repository.

The open lexical infrastructure of Språkbanken

Lars Borin, Markus Forsberg, Leif-Jöran Olsson and Jonatan Uppström

We present our ongoing work on Karp, Språkbanken's (the Swedish Language Bank) open lexical infrastructure, which has two main functions: (1) to support the work on creating, curating, and integrating our various lexical resources; and (2) to publish daily versions of the resources, making them searchable and downloadable. An important requirement on the lexical infrastructure is also that we maintain a strong bidirectional connection to our corpus infrastructure. At the heart of the infrastructure is the SweFN++ project with the goal to create free Swedish lexical resources geared towards language technology applications. The infrastructure currently hosts 15 Swedish lexical resources, including historical ones, some of which have been created from scratch using existing free resources, both external and in-house. The resources are integrated through links to a pivot lexical resource, SALDO, a large morphological and lexical-semantic resource for modern Swedish. SALDO has been selected as the pivot partly because of its size and quality, but also because its form and sense units have been assigned persistent identifiers (PIDs) to which the lexical information in other lexical resources and in corpora are linked.

The Open Linguistics Working Group

Christian Chiarcos, Sebastian Hellmann, Sebastian Nordhoff, Steven Moran, Richard Littauer, Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek and Christian M. Meyer

This paper describes the Open Linguistics Working Group (OWLG) of the Open Knowledge Foundation (OKFN). The

OWLG is an initiative concerned with linguistic data by scholars from diverse fields, including linguistics, NLP, and information science. The primary goal of the working group is to promote the idea of open linguistic resources, to develop means for their representation and to encourage the exchange of ideas across different disciplines. This paper summarizes the progress of the working group, goals that have been identified, problems that we are going to address, and recent activities and ongoing developments. Here, we put particular emphasis on the development of a Linked Open Data (sub-)cloud of linguistic resources that is currently being pursued by several OWLG members.

GATEtoGerManC: A GATE-based Annotation Pipeline for Historical German

Silke Scheible, Richard J. Whitt, Martin Durrell and Paul Bennett

We describe a new GATE-based linguistic annotation pipeline for Early Modern German, which can be used to annotate historical texts with word tokens, sentence boundaries, lemmas, and POS tags. The pipeline is based on a customisation of the freely available ANNIE system for English (Cunningham et al., 2002), in combination with a version of the TreeTagger (Schmid, 1994) trained on gold standard Early Modern German data. The POS-tagging and lemmatisation components of the pipeline achieve an average accuracy of 89.44% and 83.16%, respectively, on unseen historical data from various genres and publication dates within the Early Modern period. We show that normalisation of spelling variation can further improve these results. With no specialised tools available for processing this particular stage of the language, this pipeline will be of particular interest to smaller, humanities-based projects wishing to add linguistic annotations to their historical data but which lack the means or resources to develop such tools themselves.

Tackling interoperability issues within UIMA workflows

Nicolas Hernandez

One of the major issues dealing with any workflow management frameworks is the components interoperability. In this paper, we are concerned with the Apache UIMA framework. We address the problem by considering separately the development of new components and the integration of existing tools. For the former objective, we propose an API to generically handle TS objects by their name using reflexivity in order to make the components TS-independent. In the latter case, we distinguish the case of aggregating heterogeneous TS-dependent UIMA components from the case of integrating non UIMA-native third party tools. We propose a mapper component to aggregate TS-dependent UIMA components. And we propose a component to wrap command lines third party tools and a set of components to connect various markup languages with the UIMA data structure. Finally, we present two situations where these solutions were effectively used: Training a POS tagger system from a treebank, and embedding an external POS tagger in a workflow. Our approch aims at providing quick development solutions.

P40 - Knowledge and Ontologies

Friday, May 25, 11:45

Chairperson: **Robert Gaizauskas** Poster Session

Knowledge-Rich Context Extraction and Ranking with KnowPipe

Anne-Kathrin Schumann

This paper presents ongoing Phd thesis work dealing with the extraction of knowledge-rich contexts from text corpora for terminographic purposes. Although notable progress in the field has been made over recent years, there is yet no methodology or integrated workflow that is able to deal with multiple, typologically different languages and different domains, and that can be handled by non-expert users. Moreover, while a lot of work has been carried out to research the KRC extraction step, the selection and further analysis of results still involves considerable manual work. In this view, the aim of this paper is two-fold. Firstly, the paper presents a ranking algorithm geared at supporting the selection of high-quality contexts once the extraction has been finished and describes ranking experiments with Russian context candidates. Secondly, it presents the KnowPipe framework for context extraction: KnowPipe aims at providing a processing environment that allows users to extract knowledge-rich contexts from text corpora in different languages using shallow and deep processing techniques. In its current state of development, KnowPipe provides facilities for preprocessing Russian and German text corpora, for pattern-based knowledge-rich context extraction from these corpora using shallow analysis as well as tools for ranking Russian context candidates.

Application of a Semantic Search Algorithm to Semi-Automatic GUI Generation

Maria Teresa Pazienza, Noemi Scarpato and Armando Stellato

The Semantic Search research field aims to query metadata and to identify relevant subgraphs. While in traditional search engines queries are composed by lists of keywords connected through boolean operators, Semantic Search instead, requires the

submission of semantic queries that are structured as a graph of concepts, entities and relations. Submission of this graph is however not trivial as while a list of keywords of interest can be provided by any user, the formulation of semantic queries is not easy as well. One of the main challenges of RDF Browsers lies in the implementation of interfaces that allow the common user to submit semantic queries by hiding their complexity. Furthermore a good semantic search algorithm is not enough to fullfil user needs, it is worthwhile to implement visualization methods which can support users in intuitively understanding why and how the results were retrieved. In this paper we present a novel solution to query RDF datasets and to browse the results of the queries in an appealing manner.

The KnowledgeStore: an Entity-Based Storage System

Roldano Cattoni, Francesco Corcoglioniti, Christian Girardi, Bernardo Magnini, Luciano Serafini and Roberto Zanoli

This paper describes the KnowledgeStore, a large-scale infrastructure for the combined storage and interlinking of multimedia resources and ontological knowledge. Information in the KnowledgeStore is organized around entities, such as persons, organizations and locations. The system allows (i) to import background knowledge about entities, in form of annotated RDF triples; (ii) to associate resources to entities by automatically recognizing, coreferring and linking mentions of named entities; and (iii) to derive new entities based on knowledge extracted from mentions. The KnowledgeStore builds on state of art technologies for language processing, including document tagging, named entity extraction and cross-document coreference. Its design provides for a tight integration of linguistic and semantic features, and eases the further processing of information by explicitly representing the contexts where knowledge and mentions are valid or relevant. We describe the system and report about the creation of a large-scale KnowledgeStore instance for storing and integrating multimedia contents and background knowledge relevant to the Italian Trentino region.

Tools for plWordNet Development. Presentation and Perspectives

Bartosz Broda, Marek Maziarz and Maciej Piasecki

Building a wordnet is a serious undertaking. Fortunately, Language Technology (LT) can improve the process of wordnet construction both in terms of quality and cost. In this paper we present LT tools used during the construction of plWordNet and their influence on the lexicographer's work-flow. LT is employed in plWordNet development on every possible step: from data gathering through data analysis to data presentation. Nevertheless, every decision requires input from the lexicographer, but the quality of supporting tools is an important factor. Thus a limited evaluation of usefulness of employed tools is carried out on the basis of questionnaires.

Combining Formal Concept Analysis and semantic information for building ontological structures from texts : an exploratory study

Silvia Moraes and Vera Lima

This work studies conceptual structures based on the Formal Concept Analysis method. We build these structures based on lexico-semantic information extracted from texts, among which we highlight the semantic roles. In our research, we propose ways to include semantic roles in concepts produced by this formal method. We analyze the contribution of semantic roles and verb classes in the composition of these concepts through structural measures. In these studies, we use the Penn Treebank Sample and SemLink 1.1 corpora, both in English.

RELcat: a Relation Registry for ISOcat data categories

Menzo Windhouwer

The ISOcat Data Category Registry contains basically a flat and easily extensible list of data category specifications. To foster reuse and standardization only very shallow relationships among data categories are stored in the registry. However, to assist crosswalks, possibly based on personal views, between various (application) domains and to overcome possible proliferation of data categories more types of ontological relationships need to be specified. RELcat is a first prototype of a Relation Registry, which allows storing arbitrary relationships. These relationships can reflect the personal view of one linguist or a larger community. The basis of the registry is a relation type taxonomy that can easily be extended. This allows on one hand to load existing sets of relations specified in, for example, an OWL (2) ontology or SKOS taxonomy. And on the other hand allows algorithms that query the registry to traverse the stored semantic network to remain ignorant of the original source vocabulary. This paper describes first experiences with RELcat and explains some initial design decisions.

A disambiguation resource extracted from Wikipedia for semantic annotation

Eric Charton and Michel Gagnon

The Semantic Annotation (SA) task consists in establishing the relation between a textual entity (word or group of words designating a named entity of the real world or a concept) and

its corresponding entity in an ontology. The main difficulty of this task is that a textual entity might be highly polysemic and potentially related to many different ontological representations. To solve this specific problem, various Information Retrieval techniques can be used. Most of those involves contextual words to estimate wich exact textual entity have to be recognized. In this paper, we present a resource of contextual words that can be used by IR algorithms to establish a link between a named entity (NE) in a text and an entry point to its semantic description in the LinkedData Network.

NLP Challenges for Eunomos a Tool to Build and Manage Legal Knowledge

Guido Boella, Luigi di Caro, Llio Humphreys, Livio Robaldo and Leon van der Torre

In this paper, we describe how NLP can semi-automate the construction and analysis of knowledge in Eunomos, a legal knowledge management service which enables users to view legislation from various sources and find the right definitions and explanations of legal concepts in a given context. NLP can semi-automate some routine tasks currently performed by knowledge engineers, such as classifying norm, or linking key terms within legislation to ontological concepts. This helps overcome the resource bottleneck problem of creating specialist knowledge management systems. While accuracy is of the utmost importance in the legal domain, and the information should be verified by domain experts as a matter of course, a semi-automated approach can result in considerable efficiency gains.

Representing General Relational Knowledge in ConceptNet 5

Robert Speer and Catherine Havasi

ConceptNet is a knowledge representation project, providing a large semantic graph that describes general human knowledge and how it is expressed in natural language. This paper presents the latest iteration, ConceptNet 5, including its fundamental design decisions, ways to use it, and evaluations of its coverage and accuracy.

P41 - Semantics

Friday, May 25, 11:45

Chairperson: **Marc Verhagen** Poster Session

A new dynamic approach for lexical networks evaluation

Alain Joubert and Mathieu Lafourcade

Since September 2007, a large scale lexical network for French is under construction with methods based on popular consensus by means of games (under the JeuxDeMots project). To assess the resource quality, we decided to adopt an approach similar to its construction, that is to say an evaluation by laymen on open class vocabulary with a Tip of the Tongue tool.

LIE: Leadership, Influence and Expertise

Roberta Catizone, Louise Guthrie, Arthur Thomas and Yorick Wilks

This paper describes our research into methods for inferring social and instrumental roles and relationships from document and discourse corpora. The goal is to identify the roles of initial authors and participants in internet discussions with respect to leadership, influence and expertise. Web documents, forums and blogs provide data from which the relationships between these concepts are empirically derived and compared. Using techniques from Natural Language Processing (NLP), characterizations of authority and expertise are hypothesized and then tested to see if these pick out the same or different participants as may be chosen by techniques based on social network analysis (Huffaker 2010) see if they pick out the same discourse participants for any given level of these qualities (i.e. leadership, expertise and influence). Our methods could be applied, in principle, to any domain topic, but this paper will describe an initial investigation into two subject areas where a range of differing opinions are available and which differ in the nature of their appeals to authority and truth: ‘genetic engineering’ and a ‘Muslim Forum’. The available online corpora for these topics contain discussions from a variety of users with different levels of expertise, backgrounds and personalities.

Semantic Role Labeling with the Swedish FrameNet

Richard Johansson, Karin Friberg Heppin and Dimitrios Kokkinakis

We present the first results on semantic role labeling using the Swedish FrameNet, which is a lexical resource currently in development. Several aspects of the task are investigated, including the %design and selection of machine learning features, the effect of choice of syntactic parser, and the ability of the system to generalize to new frames and new genres. In addition, we evaluate two methods to make the role label classifier more robust: cross-frame generalization and cluster-based features. Although the small amount of training data limits the performance achievable at the moment, we reach promising results. In particular, the classifier that extracts the boundaries of arguments works well for new frames, which suggests that it already at this stage can be useful in a semi-automatic setting.

Extending a wordnet framework for simplicity and scalability

Pedro Fialho, Sérgio Curto, Ana Cristina Mendes and Luísa Coheur

The WordNet knowledge model is currently implemented in multiple software frameworks providing procedural access to language instances of it. Frameworks tend to be focused on structural/design aspects of the model thus describing low level interfaces for linguistic knowledge retrieval. Typically the only high level feature directly accessible is word lookup while traversal of semantic relations leads to verbose/complex combinations of data structures, pointers and indexes which are irrelevant in an NLP context. Here is described an extension to the JWNL framework that hides technical requirements of access to WordNet features with an essentially word/sense based API applying terminology from the official online interface. This high level API is applied to the original English version of WordNet and to an SQL based Portuguese lexicon, translated into a WordNet based representation usable by JWNL.

German “nach"-Particle Verbs in Semantic Theory and Corpus Data

Boris Haselbach, Wolfgang Seeker and Kerstin Eckart

In this paper, we present a database-supported corpus study where we combine automatically obtained linguistic information from a statistical dependency parser, namely the occurrence of a dative argument, with predictions from a theory on the argument structure of German particle verbs with“nach". The theory predicts five readings of“nach”which behave differently with respect to dative licensing in their argument structure. From a huge German web corpus, we extracted sentences for a subset of“nach"-particle verbs for which no dative is expected by the theory. Making use of a relational database management system, we bring together the corpus sentences and the lemmas manually annotated along the lines of the theory. We validate the theoretical predictions against the syntactic structure of the corpus sentences, which we obtained from a statistical dependency parser. We find that, in principle, the theory is borne out by the data, however, manual error analysis reveals cases for which the theory needs to be extended.

LexIt: A Computational Resource on Italian Argument Structure

Alessandro Lenci, Gabriella Lapesa and Giulia Bonansinga

The aim of this paper is to introduce LexIt, a computational framework for the automatic acquisition and exploration of distributional information about Italian verbs, nouns and adjectives, freely available through a web interface at the address http://sesia.humnet.unipi.it/lexit. LexIt is the first large-scale resource for Italian in which subcategorization and semantic selection properties are characterized fully on distributional ground: in the paper we describe both the process of data extraction and the evaluation of the subcategorization frames extracted with LexIt.

Enriching the ISST-TANL Corpus with Semantic Frames

Alessandro Lenci, Simonetta Montemagni, Giulia Venturi and Maria Grazia Cutrullà

The paper describes the design and the results of a manual annotation methodology devoted to enrich the ISST–TANL Corpus, derived from the Italian Syntactic–Semantic Treebank (ISST), with Semantic Frames information. The main issues encountered in applying the English FrameNet annotation criteria to a corpus of Italian language are discussed together with the choice of anchoring the semantic annotation layer to the underlying dependency syntactic structure. The results of a case study aimed at extending and specialising this methodology for the annotation of a corpus of legislative texts are also discussed.

P42 - Temporal Information

Friday, May 25, 11:45

Chairperson: **Uwe Quasthoff** Poster Session

TimeBankPT: A TimeML Annotated Corpus of Portuguese

Francisco Costa and António Branco

In this paper, we introduce TimeBankPT, a TimeML annotated corpus of Portuguese. It has been produced by adapting an existing resource for English, namely the data used in the first TempEval challenge. TimeBankPT is the first corpus of Portuguese with rich temporal annotations (i.e. it includes annotations not only of temporal expressions but also about events and temporal relations). In addition, it was subjected to an automated error mining procedure that checks the consistency of the annotated temporal relations based on their logical properties. This procedure allowed for the detection of some errors in the annotations, that also affect the original English corpus. The Portuguese language is currently undergoing a spelling reform, and several countries where Portuguese is official are in a transitional period where old and new orthographies are valid. TimeBankPT adopts the recent spelling reform. This decision is to preserve its future usefulness. TimeBankPT is freely available for download.

SUTime: A library for recognizing and normalizing time expressions

Angel X. Chang and Christopher Manning

We describe SUTIME, a temporal tagger for recognizing and normalizing temporal expressions in English text. SUTIME is available as part of the Stanford CoreNLP pipeline and can be used to annotate documents with temporal information. It is a deterministic rule-based system designed for extensibility. Testing on the TempEval-2 evaluation corpus shows that this system outperforms state-of-the-art techniques.

Temporal Annotation: A Proposal for Guidelines and an Experiment with Inter-annotator Agreement

André Bittar, Caroline Hagège, Véronique Moriceau, Xavier Tannier and Charles Teissèdre

This article presents work carried out within the framework of the ongoing ANR (French National Research Agency) project Chronolines, which focuses on the temporal processing of large news-wire corpora in English and French. The aim of the project is to create new and innovative interfaces for visualizing textual content according to temporal criteria. Extracting and normalizing the temporal information in texts through linguistic annotation is an essential step towards attaining this objective. With this goal in mind, we developed a set of guidelines for the annotation of temporal and event expressions that is intended to be compatible with the TimeML markup language, while addressing some of its pitfalls. We provide results of an initial application of these guidelines to real news-wire texts in French over several iterations of the annotation process. These results include inter-annotator agreement figures and an error analysis. Our final inter-annotator agreement figures compare favorably with those reported for the TimeBank 1.2 annotation project.

Temporal Tagging on Different Domains: Challenges, Strategies, and Gold Standards

Jannik Strötgen and Michael Gertz

In the last years, temporal tagging has received increasing attention in the area of natural language processing. However, most of the research so far concentrated on processing news documents. Only recently, two temporal annotated corpora of narrative-style documents were developed, and it was shown that a domain shift results in significant challenges for temporal tagging. Thus, a temporal tagger should be aware of the domain associated with documents that are to be processed and apply domain-specific strategies for extracting and normalizing temporal expressions. In this paper, we analyze the characteristics of temporal expressions in different domains. In addition to news- and narrative-style documents, we add two further document types, namely colloquial and scientific documents. After discussing the challenges of temporal tagging on the different domains, we describe some strategies to tackle these challenges and describe their integration into our publicly available temporal tagger HeidelTime. Our cross-domain evaluation validates the benefits of domain-sensitive temporal tagging. Furthermore, we make available two new temporally annotated corpora and a new version of HeidelTime, which now distinguishes between four document domain types.

Massively Increasing TIMEX3 Resources: A Transduction Approach

Leon Derczynski, Hector Llorens and Estela Saquete

Automatic annotation of temporal expressions is a research challenge of great interest in the field of information extraction. Gold standard temporally-annotated resources are limited in size, which makes research using them difficult. Standards have also evolved over the past decade, so not all temporally annotated data is in the same format. We vastly increase available human-annotated temporal expression resources by converting older format resources to TimeML/TIMEX3. This task is difficult due to differing annotation methods. We present a robust conversion tool and a new, large temporal expression resource. Using this, we evaluate our conversion process by using it as training data for an existing TimeML annotation tool, achieving a 0.87 F1 measure - better than any system in the TempEval-2 timex recognition exercise.

Romanian TimeBank: An Annotated Parallel Corpus for Temporal Information

Corina Forascu and Dan Tufiş

The paper describes the main steps for the construction, annotation and validation of the Romanian version of the TimeBank corpus. Starting from the English TimeBank corpus – the reference annotated corpus in the temporal domain, we have translated all the 183 English news texts into Romanian and mapped the English annotations onto Romanian, with a success rate of 96.53%. Based on ISO-Time - the emerging standard for representing temporal information, which includes many of the previous annotations schemes -, we have evaluated the automatic transfer onto Romanian and, and, when necessary, corrected the Romanian annotations so that in the end we obtained a 99.18% transfer rate for the TimeML annotations. In very few cases, due to language peculiarities, some original annotations could not be transferred. For the portability of the temporal annotation standard to Romanian, we suggested some additions for the ISO-Time standard, concerning especially the EVENT tag, based on

linguistic evidence, the Romanian grammar, and also on the localisations of TimeML to other Romance languages. Future improvements to the Ro-TimeBank will take into consideration all temporal expressions, signals and events in texts, even those with a not very clear temporal anchoring.

P43 - Sign Language

Friday, May 25, 11:45

Chairperson: **Thomas Hanke** Poster Session

Detecting Reduplication in Videos of American Sign Language

Zoya Gavrilov, Stan Sclaroff, Carol Neidle and Sven Dickinson

A framework is proposed for the detection of reduplication in digital videos of American Sign Language (ASL). In ASL, reduplication is used for a variety of linguistic purposes, including overt marking of plurality on nouns, aspectual inflection on verbs, and nominalization of verbal forms. Reduplication involves the repetition, often partial, of the articulation of a sign. In this paper, the apriori algorithm for mining frequent patterns in data streams is adapted for finding reduplication in videos of ASL. The proposed algorithm can account for varying weights on items in the apriori algorithm's input sequence. In addition, the apriori algorithm is extended to allow for inexact matching of similar hand motion subsequences and to provide robustness to noise. The formulation is evaluated on 105 lexical signs produced by two native signers. To demonstrate the formulation, overall hand motion direction and magnitude are considered; however, the formulation should be amenable to combining these features with others, such as hand shape, orientation, and place of articulation.

BiBiKit - A Bilingual Bimodal Reading and Writing Tool for Sign Language Users

Nedelina Ivanova and Olle Eriksen

Sign language is used by many people who were born deaf or who became deaf early in life use as their first and/or preferred language. There is no writing system for sign languages; texts are signed on video. As a consequence, texts in sign language are hard to navigate, search and annotate. The BiBiKit project is an easy to use authoring kit which is being developed and enables students, teachers, and virtually everyone to write and read bilingual bimodal texts and thereby creating electronic productions, which link text to sign language video. The main purpose of the project is to develop software that enables the user to link text to video, at the word, phrase and/or sentence level. The software will be developed for sign language and vice versa, but can be used to easily link text to any video: e.g. to add annotations, captions, or navigation points. The three guiding principles are: Software that is 1) stable, 2) easy to use, and 3) foolproof. A web based platform will be developed so the software is available whenever and wherever.

Resource production of written forms of Sign Languages by a user-centered editor, SWift (SignWriting improved fast transcriber)

Fabrizio Borgia, Claudia S. Bianchini, Patrice Dalle and Maria De Marsico

The SignWriting improved fast transcriber (SWift), presented in this paper, is an advanced editor for computer-aided writing and transcribing of any Sign Language (SL) using the SignWriting (SW). The application is an editor which allows composing and saving desired signs using the SW elementary components, called "glyphs". These make up a sort of alphabet, which does not depend on the national Sign Language and which codes the basic components of any sign. The user is guided through a fully automated procedure making the composition process fast and intuitive. SWift pursues the goal of helping to break down the "electronic" barriers that keep deaf people away from the web, and at the same time to support linguistic research about Sign Languages features. For this reason it has been designed with a special attention to deaf user needs, and to general usability issues. The editor has been developed in a modular way, so it can be integrated everywhere the use of the SW as an alternative to written "verbal" language may be advisable.

RWTH-PHOENIX-Weather: A Large Vocabulary Sign Language Recognition and Translation Corpus

Jens Forster, Christoph Schmidt, Thomas Hoyoux, Oscar Koller, Uwe Zelle, Justus Piater and Hermann Ney

This paper introduces the RWTH-PHOENIX-Weather corpus, a video-based, large vocabulary corpus of German Sign Language suitable for statistical sign language recognition and translation. In contrastto most available sign language data collections, the RWTH-PHOENIX-Weather corpus has not been recorded for linguistic research but for the use in statistical pattern recognition. The corpus contains weather forecasts recorded from German public TV which are manually annotated using glosses distinguishing sign variants, and time boundaries have been marked on the sentence and the gloss level. Further, the spoken German weather forecast has been transcribed in a semi-automatic fashion using a state-of-the-art automatic speech recognition system. Moreover, an additional translation of the glosses into

spoken German has been created to capture allowable translation variability. In addition to the corpus, experimental baseline results for hand and head tracking, statistical sign language recognition and translation are presented.

O45 - New Media (Special Session)

Friday, May 25, 14:55

Chairperson: **Thierry Declerk** Oral Session

Two Database Resources for Processing Social Media English Text

Eleanor Clark and Kenji Araki

This research focuses on text processing in the sphere of English-language social media. We introduce two database resources. The first, CECS (Casual English Conversion System) database, a lexicon-type resource of 1,255 entries, was constructed for use in our experimental system for the automated normalization of casual, irregularly-formed English used in communications such as Twitter. Our rule-based approach primarily aims to avoid problems caused by user creativity and individuality of language when Twitter-style text is used as input in Machine Translation, and to aid comprehension for non-native speakers of English. Although the database is still under development, we have so far carried out two evaluation experiments using our system which have shown positive results. The second database, CEGS (Casual English Generation System) phoneme database contains sets of alternative spellings for the phonemes in the CMU Pronouncing Dictionary, designed for use in a system for generating phoneme-based casual English text from regular English input; in other words, automatically producing humanlike creative sentences as an AI task. This paper provides an overview of the necessity, method, application and evaluation of both resources.

Holaaa!! writin like u talk is kewl but kinda hard 4 NLP

Maite Melero, Marta R. Costa-Jussà, Judith Domingo, Montse Marquina and Martí Quixal

We present work in progress aiming to build tools for the normalization of User-Generated Content (UGC). As we will see, the task requires the revisiting of the initial steps of NLP processing, since UGC (micro-blog, blog, and, generally, Web 2.0 user texts) presents a number of non-standard communicative and linguistic characteristics, and is in fact much closer to oral and colloquial language than to edited text. We present and characterize a corpus of UGC text in Spanish from three different sources: Twitter, consumer reviews and blogs. We motivate the need for UGC text normalization by analyzing the problems found when processing this type of text through a conventional language processing pipeline, particularly in the tasks of lemmatization and morphosyntactic tagging, and finally we propose a strategy for automatically normalizing UGC using a selector of correct forms on top of a pre-existing spell-checker.

Foundations of a Multilayer Annotation Framework for Twitter Communications During Crisis Events

William J. Corvey, Sudha Verma, Sarah Vieweg, Martha Palmer and James H. Martin

In times of mass emergency, vast amounts of data are generated via computer-mediated communication (CMC) that are difficult to manually collect and organize into a coherent picture. Yet valuable information is broadcast, and can provide useful insight into time- and safety-critical situations if captured and analyzed efficiently and effectively. We describe a natural language processing component of the EPIC (Empowering the Public with Information in Crisis) Project infrastructure, designed to extract linguistic and behavioral information from tweet text to aid in the task of information integration. The system incorporates linguistic annotation, in the form of Named Entity Tagging, as well as behavioral annotations to capture tweets contributing to situational awareness and analyze the information type of the tweet content. We show classification results and describe future integration of these classifiers in the larger EPIC infrastructure.

EmpaTweet: Annotating and Detecting Emotions on Twitter

Kirk Roberts, Michael A. Roach, Joseph Johnson, Josh Guthrie and Sanda M. Harabagiu

The rise of micro-blogging in recent years has resulted in significant access to emotion-laden text. Unlike emotion expressed in other textual sources (e.g., blogs, quotes in newswire, email, product reviews, or even clinical text), micro-blogs differ by (1) placing a strict limit on length, resulting radically in new forms of emotional expression, and (2) encouraging users to express their daily thoughts in real-time, often resulting in far more emotion statements than might normally occur. In this paper, we introduce a corpus collected from Twitter with annotated micro-blog posts (or "tweets") annotated at the tweet-level with seven emotions: ANGER, DISGUST, FEAR, JOY, LOVE, SADNESS, and SURPRISE. We analyze how emotions are distributed in the data we annotated and compare it to the distributions in other emotion-annotated corpora. We also used the annotated corpus to train a classifier that automatically discovers the emotions in tweets. In addition, we present an analysis of the linguistic style used for expressing emotions our corpus. We

hope that these observations will lead to the design of novel emotion detection techniques that account for linguistic style and psycholinguistic theories.

O46 - Semantics, Knowledge and Ontologies

Friday, May 25, 14:55

Chairperson: **Christian Chiarcos** Oral Session

Semantic Relations Established by Specialized Processes Expressed by Nouns and Verbs: Identification in a Corpus by means of Syntactico-semantic Annotation

Nava Maroto, Marie-Claude L'Homme and Amparo Alcina

This article presents the methodology and results of the analysis of terms referring to processes expressed by verbs or nouns in a corpus of specialized texts dealing with ceramics. Both noun and verb terms are explored in context in order to identify and represent the semantic roles held by their participants (arguments and circumstants), and therefore explore some of the relations established by these terms. We present a methodology for the identification of related terms that take part in the development of specialized processes and the annotation of the semantic roles expressed in these contexts. The analysis has allowed us to identify participants in the process, some of which were already present in our previous work, but also some new ones. This method is useful in the distinction of different meanings of the same verb. Contexts in which processes are expressed by verbs have proved to be very informative, even if they are less frequent in the corpus. This work is viewed as a first step in the implementation – in ontologies – of conceptual relations in which activities are involved.

Using Wikipedia to Validate the Terminology found in a Corpus of Basic Textbooks

Jorge Vivaldi, Luis Adrián Cabrera-Diego, Gerardo Sierra and María Pozzi

A scientific vocabulary is a set of terms that designate scientific concepts. This set of lexical units can be used in several applications ranging from the development of terminological dictionaries and machine translation systems to the development of lexical databases and beyond. Even though automatic term recognition systems exist since the 80s, this process is still mainly done by hand, since it generally yields more accurate results, although not in less time and at a higher cost. Some of the reasons for this are the fairly low precision and recall results obtained, the domain dependence of existing tools and the lack of available semantic knowledge needed to validate these results. In this paper we present a method that uses Wikipedia as a semantic knowledge resource, to validate term candidates from a set of scientific text books used in the last three years of high school for mathematics, health education and ecology. The proposed method may be applied to any domain or language (assuming there is a minimal coverage by Wikipedia).

PEARL: ProjEction of Annotations Rule Language, a Language for Projecting (UIMA) Annotations over RDF Knowledge Bases

Maria Teresa Pazienza, Armando Stellato and Andrea Turbati

In this paper we present a language, PEARL, for projecting annotations based on the Unstructured Information Management Architecture (UIMA) over RDF triples. The language offer is twofold: first, a query mechanism, built upon (and extending) the basic FeaturePath notation of UIMA, allows for efficient access to the standard annotation format of UIMA based on feature structures. PEARL then provides a syntax for projecting the retrieved information onto an RDF Dataset, by using a combination of a SPARQL-like notation for matching pre-existing elements of the dataset and of meta-graph patterns, for storing new information into it. In this paper we present the basics of this language and how a PEARL document is structured, discuss a simple use-case and introduce a wider project about automatic acquisition of knowledge, in which PEARL plays a pivotal role.

Constructing Large Proposition Databases

Peter Exner and Pierre Nugues

With the advent of massive online encyclopedic corpora such as Wikipedia, it has become possible to apply a systematic analysis to a wide range of documents covering a significant part of human knowledge. Using semantic parsers, it has become possible to extract such knowledge in the form of propositions (predicate–argument structures) and build large proposition databases from these documents. This paper describes the creation of multilingual proposition databases using generic semantic dependency parsing. Using Wikipedia, we extracted, processed, clustered, and evaluated a large number of propositions. We built an architecture to provide a complete pipeline dealing with the input of text, extraction of knowledge, storage, and presentation of the resulting propositions.

Highlighting relevant concepts from Topic Signatures

Montse Cuadros, Lluís Padró and German Rigau

This paper presents deepKnowNet, a new fully automatic method for building highly dense and accurate knowledge bases from

existing semantic resources. Basically, the method applies a knowledge-based Word Sense Disambiguation algorithm to assign the most appropriate WordNet sense to large sets of topically related words acquired from the web, named TSWEB. This Word Sense Disambiguation algorithm is the personalized PageRank algorithm implemented in UKB. This new method improves by automatic means the current content of WordNet by creating large volumes of new and accurate semantic relations between synsets. KnowNet was our first attempt towards the acquisition of large volumes of semantic relations. However, KnowNet had some limitations that have been overcomed with deepKnowNet. deepKnowNet disambiguates the first hundred words of all Topic Signatures from the web (TSWEB). In this case, the method highlights the most relevant word senses of each Topic Signature and filter out the ones that are not so related to the topic. In fact, the knowledge it contains outperforms any other resource when is empirically evaluated in a common framework based on a similarity task annotated with human judgements.

O47 - Segmentation,Tagging, Parsing

Friday, May 25, 14:55

Chairperson: **Valia Kordoni** Oral Session

Towards an LFG parser for Polish: An exercise in parasitic grammar development

Agnieszka Patejuk and Adam Przepiórkowski

While it is possible to build a formal grammar manually from scratch or, going to another extreme, to derive it automatically from a treebank, the development of the LFG grammar of Polish presented in this paper is different from both of these methods as it relies on extensive reuse of existing language resources for Polish. LFG grammars minimally provide two levels of representation: constituent structure (c-structure) produced by context-free phrase structure rules and functional structure (f-structure) created by functional descriptions. The c-structure was based on a DCG grammar of Polish, while the f-structure level was mainly inspired by the available HPSG analyses of Polish. The morphosyntactic information needed to create a lexicon may be taken from one of the following resources: a morphological analyser, a treebank or a corpus. Valence information from the dictionary which accompanies the DCG grammar was converted so that subcategorisation is stated in terms of grammatical functions rather than categories; additionally, missing valence frames may be extracted from the treebank. The obtained grammar is evaluated using constructed testsuites (half of which were provided by previous grammars) and the treebank.

Using a Goodness Measurement for Domain Adaptation: A Case Study on Chinese Word Segmentation

Yan Song and Fei Xia

Domain adaptation is an important topic for natural language processing. There has been extensive research on the topic and various methods have been explored, including training data selection, model combination, semi-supervised learning. In this study, we propose to use a goodness measure, namely, description length gain (DLG), for domain adaptation for Chinese word segmentation. We demonstrate that DLG can help domain adaptation in two ways: as additional features for supervised segmenters to improve system performance, and also as a similarity measure for selecting training data to better match a test set. We evaluated our systems on the Chinese Penn Treebank version 7.0, which has 1.2 million words from five different genres, and the Chinese Word Segmentation Bakeoff-3 data.

The Dependency-Parsed FrameNet Corpus

Daniel Bauer, Hagen Fürstenau and Owen Rambow

When training semantic role labeling systems, the syntax of example sentences is of particular importance. Unfortunately, for the FrameNet annotated sentences, there is no standard parsed version. The integration of the automatic parse of an annotated sentence with its semantic annotation, while conceptually straightforward, is complex in practice. We present a standard dataset that is publicly available and that can be used in future research. This dataset contains parser-generated dependency structures (with POS tags and lemmas) for all FrameNet 1.5 sentences, with nodes automatically associated with FrameNet annotations.

Predicting Phrase Breaks in Classical and Modern Standard Arabic Text

Majdi Sawalha, Claire Brierley and Eric Atwell

We train and test two probabilistic taggers for Arabic phrase break prediction on a purpose-built, “gold standard”, boundary-annotated and PoS-tagged Qur’an corpus of 77430 words and 8230 sentences. In a related LREC paper (Brierley et al., 2012), we cover dataset build. Here we report on comparative experiments with off-the-shelf N-gram and HMM taggers and coarse-grained feature sets for syntax and prosody, where the task is to predict boundary locations in an unseen test set stripped of boundary annotations by classifying words as breaks or non-breaks. The preponderance of non-breaks in the training data sets a challenging baseline success rate: 85.56%. However, we achieve significant gains in accuracy with the trigram tagger,

and significant gains in performance recognition of minority class instances with both taggers via Balanced Classification Rate. This is initial work on a long-term research project to produce annotation schemes, language resources, algorithms, and applications for Classical and Modern Standard Arabic.

Parsing Any Domain English text to CoNLL dependencies

Sudheer Kolachina and Prasanth Kolachina

It is well known that accuracies of statistical parsers trained over Penn Treebank on test sets drawn from the same corpus tend to be overestimates of their actual parsing performance. This gives rise to the need for evaluation of parsing performance on corpora from different domains. Evaluating multiple parsers on test sets from different domains can give a detailed picture about the relative strengths/weaknesses of different parsing approaches. Such information is also necessary to guide choice of parser in applications such as machine translation where text from multiple domains needs to be handled. In this paper, we report a benchmarking study of different state-of-art parsers for English, both constituency and dependency. The constituency parser output is converted into CoNLL-style dependency trees so that parsing performance can be compared across formalisms. Specifically, we train rerankers for Berkeley and Stanford parsers to study the usefulness of reranking for handling texts from different domains. The results of our experiments lead to interesting insights about the out-of-domain performance of different English parsers.

O48 - Named Entities and Subjectivity

Friday, May 25, 14:55

Chairperson: **Dan Tufis** Oral Session

Iterative Refinement and Quality Checking of Annotation Guidelines — How to Deal Effectively with Semantically Sloppy Named Entity Types, such as Pathological Phenomena

Udo Hahn, Elena Beisswanger, Ekaterina Buyko, Erik Faessler, Jenny Traumüller, Susann Schröder and Kerstin Hornbostel

We here discuss a methodology for dealing with the annotation of semantically hard to delineate, i.e., sloppy, named entity types. To illustrate sloppiness of entities, we treat an example from the medical domain, namely pathological phenomena. Based on our experience with iterative guideline refinement we propose to carefully characterize the thematic scope of the annotation by positive and negative coding lists and allow for alternative, short vs. long mention span annotations. Short spans account for canonical entity mentions (e.g., standardized disease names), while long spans cover descriptive text snippets which contain entity-specific elaborations (e.g., anatomical locations, observational details, etc.). Using this stratified approach, evidence for increasing annotation performance is provided by kappa-based inter-annotator agreement measurements over several, iterative annotation rounds using continuously refined guidelines. The latter reflects the increasing understanding of the sloppy entity class both from the perspective of guideline writers and users (annotators). Given our data, we have gathered evidence that we can deal with sloppiness in a controlled manner and expect inter-annotator agreement values around 80% for PathoJen, the pathological phenomena corpus currently under development in our lab.

GerNED: A German Corpus for Named Entity Disambiguation

Danuta Ploch, Leonhard Hennig, Angelina Duka, Ernesto William De Luca and Sahin Albayrak

Determining the real-world referents for name mentions of persons, organizations and other named entities in texts has become an important task in many information retrieval scenarios and is referred to as Named Entity Disambiguation (NED). While comprehensive datasets support the development and evaluation of NED approaches for English, there are no public datasets to assess NED systems for other languages, such as German. This paper describes the construction of an NED dataset based on a large corpus of German news articles. The dataset is closely modeled on the datasets used for the Knowledge Base Population tasks of the Text Analysis Conference, and contains gold standard annotations for the NED tasks of Entity Linking, NIL Detection and NIL Clustering. We also present first experimental results on the new dataset for each of these tasks in order to establish a baseline for future research efforts.

Centroids: Gold standards with distributional variation

Ian Lewin, Şenay Kafkas and Dietrich Rebholz-Schuhmann

Motivation: Gold Standards for named entities are, ironically, not standard themselves. Some specify the “one perfect annotation”. Others specify “perfectly good alternatives”. The concept of Silver standard is relatively new. The objective is consensus rather than perfection. How should the two concepts be best represented and related? Approach: We examine several Biomedical Gold Standards and motivate a new representational format, centroids, which simply and effectively represents name distributions. We define an algorithm for finding centroids, given a set of alternative input annotations and we test the outputs quantitatively and

qualitatively. We also define a metric of relatively acceptability on top of the centroid standard. Results: Precision, recall and F-scores of over 0.99 are achieved for the simple sanity check of giving the algorithm Gold Standard inputs. Qualitative analysis of the differences very often reveals errors and incompleteness in the original Gold Standard. Given automatically generated annotations, the centroids effectively represent the range of those contributions and the quality of the centroid annotations is highly competitive with the best of the contributors. Conclusion: Centroids cleanly represent alternative name variations for Silver and Gold Standards. A centroid Silver Standard is derived just like a Gold Standard, only from imperfect inputs.

Quantising Opinions for Political Tweets Analysis

Yulan He, Hassan Saif, Zhongyu Wei and Kam-Fai Wong

There have been increasing interests in recent years in analyzing tweet messages relevant to political events so as to understand public opinions towards certain political issues. We analyzed tweet messages crawled during the eight weeks leading to the UK General Election in May 2010 and found that activities at Twitter is not necessarily a good predictor of popularity of political parties. We then proceed to propose a statistical model for sentiment detection with side information such as emoticons and hash tags implying tweet polarities being incorporated. Our results show that sentiment analysis based on a simple keyword matching against a sentiment lexicon or a supervised classifier trained with distant supervision does not correlate well with the actual election results. However, using our proposed statistical model for sentiment analysis, we were able to map the public opinion in Twitter with the actual offline sentiment in real world.

AWATIF: A Multi-Genre Corpus for Modern Standard Arabic Subjectivity and Sentiment Analysis

Muhammad Abdul-Mageed and Mona Diab

We present AWATIF, a multi-genre corpus of Modern Standard Arabic (MSA) labeled for subjectivity and sentiment analysis (SSA) at the sentence level. The corpus is labeled using both regular as well as crowd sourcing methods under three different conditions with two types of annotation guidelines. We describe the sub-corpora constituting the corpus and provide examples from the various SSA categories. In the process, we present our linguistically-motivated and genre-nuanced annotation guidelines and provide evidence showing their impact on the labeling task.

P44 - Machine Translation (2)

Friday, May 25, 14:55

Chairperson: **Jan Hajic** Poster Session

Arabic-Segmentation Combination Strategies for Statistical Machine Translation

Saab Mansour and Hermann Ney

Arabic segmentation was already applied successfully for the task of statistical machine translation (SMT). Yet, there is no consistent comparison of the effect of different techniques and methods over the final translation quality. In this work, we use existing tools and further re-implement and develop new methods for segmentation. We compare the resulting SMT systems based on the different segmentation methods over the small IWSLT 2010 BTEC and the large NIST 2009 Arabic-to-English translation tasks. Our results show that for both small and large training data, segmentation yields strong improvements, but, the differences between the top ranked segmenters are statistically insignificant. Due to the different methodologies that we apply for segmentation, we expect a complimentary variation in the results achieved by each method. As done in previous work, we combine several segmentation schemes of the same model but achieve modest improvements. Next, we try a different strategy, where we combine the different segmentation methods rather than the different segmentation schemes. In this case, we achieve stronger improvements over the best single system. Finally, combining schemes and methods has another slight gain over the best combination strategy.

The Joy of Parallelism with CzEng 1.0

Ondřej Bojar, Zdeněk Žabokrtský, Ondřej Dušek, Petra Galuščáková, Martin Majliš, David Mareček, Jiří Maršík, Michal Novák, Martin Popel and Aleš Tamchyna

CzEng 1.0 is an updated release of our Czech-English parallel corpus, freely available for non-commercial research or educational purposes. In this release, we approximately doubled the corpus size, reaching 15 million sentence pairs (about 200 million tokens per language). More importantly, we carefully filtered the data to reduce the amount of non-matching sentence pairs. CzEng 1.0 is automatically aligned at the level of sentences as well as words. We provide not only the plain text representation, but also automatic morphological tags, surface syntactic as well as deep syntactic dependency parse trees and automatic co-reference links in both English and Czech. This paper describes key properties of the released resource including the distribution of text domains, the corpus data formats, and a toolkit to handle the provided rich annotation. We also summarize the procedure of the rich annotation (incl. co-reference

resolution) and of the automatic filtering. Finally, we provide some suggestions on exploiting such an automatically annotated sentence-parallel corpus.

Statistical Machine Translation without Source-side Parallel Corpus Using Word Lattice and Phrase Extension

Takanori Kusumoto and Tomoyosi Akiba

Statistical machine translation (SMT) requires a parallel corpus between the source and target languages. Although a pivot-translation approach can be applied to a language pair that does not have a parallel corpus directly between them, it requires both source–pivot and pivot–target parallel corpora. We propose a novel approach to apply SMT to a resource-limited source language that has no parallel corpus but has only a word dictionary for the pivot language. The problems with dictionary-based translations lie in their ambiguity and incompleteness. The proposed method uses a word lattice representation of the pivot-language candidates and word lattice decoding to deal with the ambiguity; the lattice expansion is accomplished by using a pivot–target phrase translation table to compensate for the incompleteness. Our experimental evaluation showed that this approach is promising for applying SMT, even when a source-side parallel corpus is lacking.

Automatic Translation of Scientific Documents in the HAL Archive

Lambert Patrik, Holger Schwenk and Frédéric Blain

This paper describes the development of a statistical machine translation system between French and English for scientific papers. This system will be closely integrated into the French HAL open archive, a collection of more than 100.000 scientific papers. We describe the creation of in-domain parallel and monolingual corpora, the development of a domain specific translation system with the created resources, and its adaptation using monolingual resources only. These techniques allowed us to improve a generic system by more than 10 BLEU points.

Expanding Parallel Resources for Medium-Density Languages for Free

Georgi Iliev and Angel Genov

We discuss a previously proposed method for augmenting parallel corpora of limited size for the purposes of machine translation through monolingual paraphrasing of the source language. We develop a three-stage shallow paraphrasing procedure to be applied to the Swedish-Bulgarian language pair for which limited parallel resources exist. The source language exhibits specifics not typical of high-density languages already studied in a similar setting. Paraphrases of a highly productive type of compound nouns in Swedish are generated by a corpus-based technique. Certain Swedish noun-phrase types are paraphrased using basic heuristics. Further we introduce noun-phrase morphological variations for better wordform coverage. We evaluate the performance of a phrase-based statistical machine translation system trained on a baseline parallel corpus and on three stages of artificial enlargement of the source-language training data. Paraphrasing is shown to have no effect on performance for the Swedish-English translation task. We show a small, yet consistent, increase in the BLEU score of Swedish-Bulgarian translations of larger token spans on the first enlargement stage. A small improvement in the overall BLEU score of Swedish-Bulgarian translation is achieved on the second enlargement stage. We find that both improvements justify further research into the method for the Swedish-Bulgarian translation task.

VERTa: Linguistic features in MT evaluation

Elisabet Comelles, Jordi Atserias, Victoria Arranz and Irene Castellón

In the last decades, a wide range of automatic metrics that use linguistic knowledge has been developed. Some of them are based on lexical information, such as METEOR; others rely on the use of syntax, either using constituent or dependency analysis; and others use semantic information, such as Named Entities and semantic roles. All these metrics work at a specific linguistic level, but some researchers have tried to combine linguistic information, either by combining several metrics following a machine-learning approach or focusing on the combination of a wide variety of metrics in a simple and straightforward way. However, little research has been conducted on how to combine linguistic features from a linguistic point of view. In this paper we present VERTa, a metric which aims at using and combining a wide variety of linguistic features at lexical, morphological, syntactic and semantic level. We provide a description of the metric and report some preliminary experiments which will help us to discuss the use and combination of certain linguistic features in order to improve the metric performance

Linguistic Resources for Handwriting Recognition and Translation Evaluation

Zhiyi Song, Safa Ismael, Stephen Grimes, David Doermann and Stephanie Strassel

We describe efforts to create corpora to support development and evaluation of handwriting recognition and translation technology. LDC has developed a stable pipeline and infrastructures for collecting and annotating handwriting linguistic resources to support the evaluation of MADCAT and OpenHaRT. We collect and annotate handwritten samples of pre-processed Arabic and

Chinese data that has been already translated in English that is used in the GALE program. To date, LDC has recruited more than 600 scribes and collected, annotated and released more than 225,000 handwriting images. Most linguistic resources created for these programs will be made available to the larger research community by publishing in LDC's catalog. The phase 1 MADCAT corpus is now available.

Development and Application of a Cross-language Document Comparability Metric

Fangzhong Su and Bogdan Babych

In this paper we present a metric that measures comparability of documents across different languages. The metric is developed within the FP7 ICT ACCURAT project, as a tool for aligning comparable corpora on the document level; further these aligned comparable documents are used for phrase alignment and extraction of translation equivalents, with the aim to extend phrase tables of statistical MT systems without the need to use parallel texts. The metric uses several features, such as lexical information, document structure, keywords and named entities, which are combined in an ensemble manner. We present the results by measuring the reliability and effectiveness of the metric, and demonstrate its application and the impact for the task of parallel phrase extraction from comparable corpora.

Assessing Divergence Measures for Automated Document Routing in an Adaptive MT System

Claire Jaja, Douglas Briesch, Jamal Laoudi and Clare Voss

Custom machine translation (MT) engines systematically outperform general-domain MT engines when translating within the relevant custom domain. This paper investigates the use of the Jensen-Shannon divergence measure for automatically routing new documents within a translation system with multiple MT engines to the appropriate custom MT engine in order to obtain the best translation. Three distinct domains are compared, and the impact of the language, size, and preprocessing of the documents on the Jensen-Shannon score is addressed. Six test datasets are then compared to the three known-domain corpora to predict which of the three custom MT engines they would be routed to at runtime given their Jensen-Shannon scores. The results are promising for incorporating this divergence measure into a translation workflow.

A Study of Word-Classing for MT Reordering

Ananthakrishnan Ramanathan and Karthik Visweswariah

MT systems typically use parsers to help reorder constituents. However most languages do not have adequate treebank data to learn good parsers, and such training data is extremely time-consuming to annotate. Our earlier work has shown that a reordering model learned from word-alignments using POS tags as features can improve MT performance (Visweswariah et al., 2011). In this paper, we investigate the effect of word-classing on reordering performance using this model. We show that unsupervised word clusters perform somewhat worse but still reasonably well, compared to a part-of-speech (POS) tagger built with a small amount of annotated data; while a richer tag set including case and gender-number-person further improves reordering performance by around 1.2 monolingual BLEU points. While annotating this richer tagset is more complicated than annotating the base tagset, it is much easier than annotating treebank data.

Dealing with unknown words in statistical machine translation

João Silva, Luísa Coheur, Ângela Costa and Isabel Trancoso

In Statistical Machine Translation, words that were not seen during training are unknown words, that is, words that the system will not know how to translate. In this paper we contribute to this research problem by profiting from orthographic cues given by words. Thus, we report a study of the impact of word distance metrics in cognates' detection and, in addition, on the possibility of obtaining possible translations of unknown words through Logical Analogy. Our approach is tested in the translation of corpora from Portuguese to English (and vice-versa).

PET: a Tool for Post-editing and Assessing Machine Translation

Wilker Aziz, Sheila Castilho and Lucia Specia

Given the significant improvements in Machine Translation (MT) quality and the increasing demand for translations, post-editing of automatic translations is becoming a popular practice in the translation industry. It has been shown to allow for much larger volumes of translations to be produced, saving time and costs. In addition, the post-editing of automatic translations can help understand problems in such translations and this can be used as feedback for researchers and developers to improve MT systems. Finally, post-editing can be used as a way of evaluating the quality of translations in terms of how much post-editing effort these translations require. We describe a standalone tool that has two main purposes: facilitate the post-editing of translations from any MT system so that they reach publishable quality and collect sentence-level information from the post-editing process, e.g.: post-editing time and detailed keystroke statistics.

Tajik-Farsi Persian Transliteration Using Statistical Machine Translation

Chris Irwin Davis

Tajik Persian is a dialect of Persian spoken primarily in Tajikistan and written with a modified Cyrillic alphabet. Iranian Persian, or Farsi, as it is natively called, is the lingua franca of Iran and is written with the Persian alphabet, a modified Arabic script. Although the spoken versions of Tajik and Farsi are mutually intelligible to educated speakers of both languages, the difference between the writing systems constitutes a barrier to text compatibility between the two languages. This paper presents a system to transliterate text between these two different Persian dialects that use incompatible writing systems. The system also serves as a mechanism to facilitate sharing of computational linguistic resources between the two languages. This is relevant because of the disparity in resources for Tajik versus Farsi.

Assessing the Comparability of News Texts

Emma Barker and Robert Gaizauskas

Comparable news texts are frequently proposed as a potential source of alignable subsentential fragments for use in statistical machine translation systems. But can we assess just how potentially useful they will be? In this paper we first discuss a scheme for classifying news text pairs according to the degree of relatedness of the events they report and investigate how robust this classification scheme is via a multi-lingual annotation exercise. We then propose an annotation methodology, similar to that used in summarization evaluation, to allow us to identify and quantify shared content at the subsentential level in news text pairs and report a preliminary exercise to assess this method. We conclude by discussing how this works fits into a broader programme of assessing the potential utility of comparable news texts for extracting paraphrases/translational equivalents for use in language processing applications.

P45 - Natural Language Generation

Friday, May 25, 14:55

Chairperson: **Dan Cristea** Poster Session

Corpus-based Referring Expressions Generation

Hilder Pereira, Eder Novais, Andre Mariotti and Ivandre Paraboni

In Natural Language Generation, the task of attribute selection (AS) consists of determining the appropriate attribute-value pairs (or semantic properties) that represent the contents of a referring expression. Existing work on AS includes a wide range of algorithmic solutions to the problem, but the recent availability of corpora annotated with referring expressions data suggests that corpus-based AS strategies become possible as well. In this work we tentatively discuss a number of AS strategies using both semantic and surface information obtained from a corpus of this kind. Relying on semantic information, we attempt to learn both global and individual AS strategies that could be applied to a standard AS algorithm in order to generate descriptions found in the corpus. As an alternative, and perhaps less traditional approach, we also use surface information to build statistical language models of the referring expressions that are most likely to occur in the corpus, and let the model probabilities guide attribute selection.

Portuguese Text Generation from Large Corpora

Eder Novais, Ivandre Paraboni and Douglas Silva

In the implementation of a surface realisation engine, many of the computational techniques seen in other AI fields have been widely applied. Among these, the use of statistical methods has been particularly successful, as in the so-called 'generate-and-select', or 2-stages architectures. Systems of this kind produce output strings from possibly underspecified input data by over-generating a large number of alternative realisations (often including ungrammatical candidate sentences.) These are subsequently ranked with the aid of a statistical language model, and the most likely candidate is selected as the output string. Statistical approaches may however face a number of difficulties. Among these, there is the issue of data sparseness, a problem that is particularly evident in cases such as our target language - Brazilian Portuguese - which is not only morphologically-rich, but relatively poor in NLP resources such as large, publicly available corpora. In this work we describe a first implementation of a shallow surface realisation system for this language that deals with the issue of data sparseness by making use of factored language models built from a (relatively) large corpus of Brazilian newspapers articles.

DSim, a Danish Parallel Corpus for Text Simplification

Sigrid Klerke and Anders Søgaard

We present DSim, a new sentence aligned Danish monolingual parallel corpus extracted from 3701 pairs of news telegrams and corresponding professionally simplified short news articles. The corpus is intended for building automatic text simplification for adult readers. We compare DSim to different examples of monolingual parallel corpora, and we argue that this corpus is a promising basis for future development of automatic data-driven text simplification systems in Danish. The corpus contains

both the collection of paired articles and a sentence aligned bitext, and we show that sentence alignment using simple tf*idf weighted cosine similarity scoring is on line with state–of–the–art when evaluated against a hand-aligned sample. The alignment results are compared to state of the art for English sentence alignment. We finally compare the source and simplified sides of the corpus in terms of lexical and syntactic characteristics and readability, and find that the one–to–many sentence aligned corpus is representative of the sentence simplifications observed in the unaligned collection of article pairs.

Acquisition of Syntactic Simplification Rules for French

Violeta Seretan

Text simplification is the process of reducing the lexical and syntactic complexity of a text while attempting to preserve (most of) its information content. It has recently emerged as an important research area, which holds promise for enhancing the text readability for the benefit of a broader audience as well as for increasing the performance of other applications. Our work focuses on syntactic complexity reduction and deals with the task of corpus-based acquisition of syntactic simplification rules for the French language. We show that the data-driven manual acquisition of simplification rules can be complemented by the semi-automatic detection of syntactic constructions requiring simplification. We provide the first comprehensive set of syntactic simplification rules for French, whose size is comparable to similar resources that exist for English and Brazilian Portuguese. Unlike these manually-built resources, our resource integrates larger lists of lexical cues signaling simplifiable constructions, that are useful for informing practical systems.

A Repository of Data and Evaluation Resources for Natural Language Generation

Anja Belz and Albert Gatt

Starting in 2007, the field of natural language generation (NLG) has organised shared-task evaluation events every year, under the Generation Challenges umbrella. In the course of these shared tasks, a wealth of data has been created, along with associated task definitions and evaluation regimes. In other contexts too, sharable NLG data is now being created. In this paper, we describe the online repository that we have created as a one-stop resource for obtaining NLG task materials, both from Generation Challenges tasks and from other sources, where the set of materials provided for each task consists of (i) task definition, (ii) input and output data, (iii) evaluation software, (iv) documentation, and (v) publications reporting previous results.

LG-Eval: A Toolkit for Creating Online Language Evaluation Experiments

Eric Kow and Anja Belz

In this paper we describe the LG-Eval toolkit for creating online language evaluation experiments. LG-Eval is the direct result of our work setting up and carrying out the human evaluation experiments in several of the Generation Challenges shared tasks. It provides tools for creating experiments with different kinds of rating tools, allocating items to evaluators, and collecting the evaluation scores.

P46 - Crowdsourcing

Friday, May 25, 14:55

Chairperson: **Collin Baker** Poster Session

Turk Bootstrap Word Sense Inventory 2.0: A Large-Scale Resource for Lexical Substitution

Chris Biemann

This paper presents the Turk Bootstrap Word Sense Inventory (TWSI) 2.0. This lexical resource, created by a crowdsourcing process using Amazon Mechanical Turk (http://www.mturk.com), encompasses a sense inventory for lexical substitution for 1,012 highly frequent English common nouns. Along with each sense, a large number of sense-annotated occurrences in context are given, as well as a weighted list of substitutions. Sense distinctions are not motivated by lexicographic considerations, but driven by substitutability: two usages belong to the same sense if their substitutions overlap considerably. After laying out the need for such a resource, the data is characterized in terms of organization and quantity. Then, we briefly describe how this data was used to create a system for lexical substitutions. Training a supervised lexical substitution system on a smaller version of the resource resulted in well over 90% acceptability for lexical substitutions provided by the system. Thus, this resource can be used to set up reliable, enabling technologies for semantic natural language processing (NLP), some of which we discuss briefly.

Collection of a Large Database of French-English SMT Output Corrections

Marion Potet, Emmanuelle Esperança-Rodier, Laurent Besacier and Hervé Blanchon

Corpus-based approaches to machine translation (MT) rely on the availability of parallel corpora. To produce user-acceptable translation outputs, such systems need high quality data to be efficiency trained, optimized and evaluated. However, building

high quality dataset is a relatively expensive task. In this paper, we describe the data collection and analysis of a large database of 10.881 SMT translation output hypotheses manually corrected. These post-editions were collected using Amazon's Mechanical Turk, following some ethical guidelines. A complete analysis of the collected data pointed out a high quality of the corrections with more than 87 % of the collected post-editions that improve hypotheses and more than 94 % of the crowdsourced post-editions which are at least of professional quality. We also post-edited 1,500 gold-standard reference translations (of bilingual parallel corpora generated by professional) and noticed that 72 % of these translations needed to be corrected during post-edition. We computed a proximity measure between the differents kind of translations and pointed out that reference translations are as far from the hypotheses than from the corrected hypotheses (i.e. the post-editions). In light of these last findings, we discuss the adequation of text-based generated reference translations to train setence-to-sentence based SMT systems.

Getting more data – Schoolkids as annotators

Jirka Hana and Barbora Hladka

We present a new way to get more morphologically and syntactically annotated data. We have developed an annotation editor tailored to school children to involve them in text annotation. Using this editor, they practice morphology and dependency-based syntax in the same way as they normally do at (Czech) schools, without any special training. Their annotation is then automatically transformed into the target annotation schema. The editor is designed to be language independent, however the subsequent transformation is driven by the annotation framework we are heading for. In our case, the object language is Czech and the target annotation scheme corresponds to the Prague Dependency Treebank annotation framework.

Word Sense Inventories by Non-Experts.

Anna Rumshisky, Nick Botchan, Sophie Kushkuley and James Pustejovsky

In this paper, we explore different strategies for implementing a crowdsourcing methodology for a single-step construction of an empirically-derived sense inventory and the corresponding sense-annotated corpus. We report on the crowdsourcing experiments using implementation strategies with different HIT costs, worker qualification testing, and other restrictions. We describe multiple adjustments required to ensure successful HIT design, given significant changes within the crowdsourcing community over the last three years.

P47 - Text Mining and Text Entailment

Friday, May 25, 14:55

Chairperson: **Sophia Ananiadou** Poster Session

The BladeMistress Corpus: From Talk to Action in Virtual Worlds

Anton Leuski, Carsten Eickhoff, James Ganis and Victor Lavrenko

Virtual Worlds (VW) are online environments where people come together to interact and perform various tasks. The chat transcripts of interactions in VWs pose unique opportunities and challenges for language analysis: Firstly, the language of the transcripts is very brief, informal, and task-oriented. Secondly, in addition to chat, a VW system records users' in-world activities. Such a record could allow us to analyze how the language of interactions is linked to the users actions. For example, we can make the language analysis of the users dialogues more effective by taking into account the context of the corresponding action or we can predict or detect users actions by analyzing the content of conversations. Thirdly, a joined analysis of both the language and the actions would empower us to build effective modes of the users and their behavior. In this paper we present a corpus constructed from logs from an online multiplayer game BladeMistress. We describe the original logs, annotations that we created on the data, and summarize some of the experiments.

Annotating Factive Verbs

Alvin Grissom II and Yusuke Miyao

We have created a scheme for annotating corpora designed to capture relevant aspects of factivity in verb-complement constructions. Factivity constructions are a well-known linguistic phenomenon that embed presuppositions about the state of the world into a clause. These embedded presuppositions provide implicit information about facts assumed to be true in the world, and are thus potentially valuable in areas of research such as textual entailment. We attempt to address both clear-cut cases of factivity and non-factivity, as well as account for the fluidity and ambiguous nature of some realizations of this construction. Our extensible scheme is designed to account for distinctions between claims, performatives, atypical uses of factivity, and the authority of the one making the utterance. We introduce a simple XML-based syntax for the annotation of factive verbs and clauses, in order to capture this information. We also provide an analysis of the issues which led to these annotative decisions, in the hope that these analyses will be beneficial to those dealing with factivity in a practical context.

A Holistic Approach to Bilingual Sentence Fragment Extraction from Comparable Corpora

Mahdi Khademian, Kaveh Taghipour, Saab Mansour and Shahram Khadivi

Achieving accurate translation, especially in multiple domain documents with statistical machine translation systems, requires more and more bilingual texts and this need becomes more critical when training such systems for language pairs with scarce training data. In the recent years, there have been some researches on new sources of parallel texts that are documents which are not necessarily parallel but are comparable. Since these methods search for possible translation equivalences in a greedy manner, they are unable to consider all possible parallel texts in comparable documents. This paper investigates a different approach for this need by considering relationships between all words of two comparable documents, which works fairly well even in the worst case of comparability. We represent each document pair in a matrix and then transform it to a new space to find parallel fragments. Evaluations show that the system is successful in extraction of useful fragment pairs.

An Examination of Cross-Cultural Similarities and Differences from Social Media Data with respect to Language Use

Mohammad Fazleh Elahi and Paola Monachesi

We present a methodology for analyzing cross-cultural similarities and differences using language as a medium, love as domain, social media as a data source and 'Terms' and 'Topics' as cultural features. We discuss the techniques necessary for the creation of the social data corpus from which emotion terms have been extracted using NLP techniques. Topics of love discussion were then extracted from the corpus by means of Latent Dirichlet Allocation (LDA). Finally, on the basis of these features, a cross-cultural comparison was carried out. For the purpose of cross-cultural analysis, the experimental focus was on comparing data from a culture from the East (India) with a culture from the West (United States of America). Similarities and differences between these cultures have been analyzed with respect to the usage of emotions, their intensities and the topics used during love discussion in social media.

Turkish Paraphrase Corpus

Seniz Demir, Ilknur Durgar El-Kahlout, Erdem Unal and Hamza Kaya

Paraphrases are alternative syntactic forms in the same language expressing the same semantic content. Speakers of all languages are inherently familiar with paraphrases at different levels of granularity (lexical, phrasal, and sentential). For quite some time, the concept of paraphrasing is getting a growing attention by the research community and its potential use in several natural language processing applications (such as text summarization and machine translation) is being investigated. In this paper, we present, what is to our best knowledge, the first Turkish paraphrase corpus. The corpus is gleaned from four different sources and currently contains 1270 paraphrase pairs. All paraphrase pairs are carefully annotated by native Turkish speakers with the identified semantic correspondences between paraphrases. The work for expanding the corpus is still under way.

Constructing a Question Corpus for Textual Semantic Relations

Rui Wang and Shuguang Li

Finding useful questions is a challenging task in Community Question Answering (CQA). There are two key issues need to be resolved: 1) what is a useful question to the given reference question; and furthermore 2) what kind of relations exist between a given pair of questions. In order to answer these two questions, in this paper, we propose a fine-grained inventory of textual semantic relations between questions and annotate a corpus constructed from the WikiAnswers website. We also extract large archives of question pairs with user-generated links and use them as labeled data for separating useful questions from neutral ones, achieving 72.2% of accuracy. We find such online CQA repositories valuable resources for related research.

Evaluating the Similarity Estimator component of the TWIN Personality-based Recommender System

Alexandra Roshchina, John Cardiff and Paolo Rosso

With the constant increase in the amount of information available in online communities, the task of building an appropriate Recommender System to support the user in her decision making process is becoming more and more challenging. In addition to the classical collaborative filtering and content based approaches, taking into account ratings, preferences and demographic characteristics of the users, a new type of Recommender System, based on personality parameters, has been emerging recently. In this paper we describe the TWIN (Tell Me What I Need) Personality Based Recommender System, and report on our experiments and experiences of utilizing techniques which allow the extraction of the personality type from text (following the Big Five model popular in the psychological research). We estimate the possibility of constructing the personality-based Recommender System that does not require users to fill in personality questionnaires. We are applying the proposed system

in the online travelling domain to perform TripAdvisor hotels recommendation by analysing the text of user generated reviews, which are freely accessible from the community website.

P48 - Speech/Multimodal Tools, Systems, Applications

Friday, May 25, 14:55

Chairperson: **Tomaž Erjavec** Poster Session

Annotation Facilities for the Reliable Analysis of Human Motion

Michael Kipp

Human motion is challenging to analyze due to the many degrees of freedom of the human body. While the qualitative analysis of human motion lies at the core of many research fields, including multimodal communication, it is still hard to achieve reliable results when human coders transcribe motion with abstract categories. In this paper we tackle this problem in two respects. First, we provide facilities for qualitative and quantitative comparison of annotations. Second, we provide facilities for exploring highly precise recordings of human motion (motion capture) using a low-cost consumer device (Kinect). We present visualization and analysis methods, integrated in the existing ANVIL video annotation tool (Kipp 2001), and provide both a precision analysis and a “cookbook” for Kinect-based motion analysis.

Translog-II: a Program for Recording User Activity Data for Empirical Reading and Writing Research

Michael Carl

This paper presents a novel implementation of Translog-II. Translog-II is a Windows-oriented program to record and study reading and writing processes on a computer. In our research, it is an instrument to acquire objective, digital data of human translation processes. As their predecessors, Translog 2000 and Translog 2006, also Translog-II consists of two main components: Translog-II Supervisor and Translog-II User, which are used to create a project file, to run a text production experiments (a user reads, writes or translates a text) and to replay the session. Translog produces a log files which contains all user activity data of the reading, writing, or translation session, and which can be evaluated by external tools. While there is a large body of translation process research based on Translog, this paper gives an overview of the Translog-II functions and its data visualization options.

Intelligibility assessment in forensic applications

Giovanni Costantini, Andrea Paoloni and Massimiliano Todisco

In the context of forensic phonetics the transcription of intercepted signals is particularly important. However, these signals are often degraded and the transcript may not reflect what was actually pronounced. In the absence of the original signal, the only way to see the level of accuracy that can be obtained in the transcription of poor recordings is to develop an objective methodology for intelligibility measurements. This study has been carried out on a corpus specially built to simulate the real conditions of forensic signals. With reference to this corpus a measurement system of intelligibility based on STI (Speech Transmission Index) has been evaluated so as to assess its performance. The result of the experiment shows a high correlation between objective measurements and subjective evaluations. Therefore it is recommended to use the proposed methodology in order to establish whether a given intercepted signal can be transcribed with sufficient reliability.

Strategies to Improve a Speaker Diarisation Tool

David Tavarez, Eva Navas, Daniel Erro and Ibon Saratxaga

This paper describes the different strategies used to improve the results obtained by an off-line speaker diarisation tool with the Albayzin 2010 diarisation database. The errors made by the system have been analyzed and different strategies have been proposed to reduce each kind of error. Very short segments incorrectly labelled and different appearances of one speaker labelled with different identifiers are the most common errors. A post-processing module that refines the segmentation by retraining the GMM models of the speakers involved has been built to cope with these errors. This post-processing module has been tuned with the training dataset and improves the result of the diarisation system by 16.4% in the test dataset.

Using an ASR database to design a pronunciation evaluation system in Basque

Igor Odriozola, Eva Navas, Inma Hernáez, Iñaki Sainz, Ibon Saratxaga, Jon Sánchez and Daniel Erro

This paper presents a method to build CAPT systems for under resourced languages, as Basque, using a general purpose ASR speech database. More precisely, the proposed method consists in automatically determine the threshold of GOP (Goodness Of Pronunciation) scores, which have been used as pronunciation scores in phone-level. Two score distributions have been obtained for each phoneme corresponding to its correct and incorrect

pronunciations. The distribution of the scores for erroneous pronunciation has been calculated inserting controlled errors in the dictionary, so that each changed phoneme has been randomly replaced by a phoneme from the same group. These groups have been obtained by means of a phonetic clustering performed using regression trees. After obtaining both distributions, the EER (Equal Error Rate) of each distribution pair has been calculated and used as a decision threshold for each phoneme. The results show that this method is useful when there is no database specifically designed for CAPT systems, although it is not as accurate as those specifically designed for this purpose.

W-PhAMT: A web tool for phonetic multilevel timeline visualization

Francesco Cutugno, Vincenza Anna Leano and Antonio Origlia

This paper presents a web platform with an its own graphic environment to visualize and filter multilevel phonetic annotations. The tool accepts as input Annotation Graph XML and Praat TextGrids files and converts these files into a specific XML format. XML output is used to browse data by means of a web tool using a visualization metaphor, namely a timeline. A timeline is a graphical representation of a period of time, on which relevant events are marked. Events are usually distributed over many layers in a geometrical metaphor represented by segments and points spatially distributed with reference to a temporal axis. The tool shows all the annotations included in the uploaded dataset, allowing the listening of the entire file or of its parts. Filtering is allowed on annotation labels by means of string pattern matching. The web service includes cloud services to share data with other users. The tool is available at http://w-phamt.fisica.unina.it

English to Indonesian Transliteration to Support English Pronunciation Practice

Amalia Zahra and Julie Carson-Berndsen

The work presented in this paper explores the use of Indonesian transliteration to support English pronunciation practice. It is mainly aimed for Indonesian speakers who have no or minimum English language skills. The approach implemented combines a rule-based and a statistical method. The rules of English-Phone-to-Indonesian-Grapheme mapping are implemented with a Finite State Transducer (FST), followed by a statistical method which is a grapheme-based trigram language model. The Indonesian transliteration generated was used as a means to support the learners where their speech were then recorded. The speech recordings have been evaluated by 19 participants: 8 English native and 11 non-native speakers. The results show that the transliteration positively contributes to the improvement of their English pronunciation.

Rapidly Testing the Interaction Model of a Pronunciation Training System via Wizard-of-Oz

Joao Paulo Cabral, Mark Kane, Zeeshan Ahmed, Mohamed Abou-Zleikha, Eva Szekely, Amalia Zahra, Kalu Ogbureke, Peter Cahill, Julie Carson-Berndsen and Stephan Schlogl

This paper describes a prototype of a computer-assisted pronunciation training system called MySpeech. The interface of the MySpeech system is web-based and it currently enables users to practice pronunciation by listening to speech spoken by native speakers and tuning their speech production to correct any mispronunciations detected by the system. This practice exercise is facilitated in different topics and difficulty levels. An experiment was conducted in this work that combines the MySpeech service with the WebWOZ Wizard-of-Oz platform (http://www.webwoz.com), in order to improve the human-computer interaction (HCI) of the service and the feedback that it provides to the user. The employed Wizard-of-Oz method enables a human (who acts as a wizard) to give feedback to the practising user, while the user is not aware that there is another person involved in the communication. This experiment permitted to quickly test an HCI model before its implementation on the MySpeech system. It also allowed to collect input data from the wizard that can be used to improve the proposed model. Another outcome of the experiment was the preliminary evaluation of the pronunciation learning service in terms of user satisfaction, which would be difficult to conduct before integrating the HCI part.

PAMOCAT: Automatic retrieval of specified postures

Bernhard Brüning, Christian Schnier, Karola Pitsch and Sven Wasmuth

In order to understand and model the non-verbal communicative conduct of humans, it seems fruitful to combine qualitative methods (Conversation Analysis) and quantitative techniques (motion capturing). A Tools for data visualization and annotation is important as they constitute a central interface between different research approaches and methodologies. We have developed the pre-annotation tool “PAMOCAT” that detects motion segments of individual joints. A sophisticated user interface easily allows the annotating person to find correlations between different joints and to export combined qualitative and quantitative annotations to standard annotation tools. Using this technique we are able to examine complex setups with three persons in tight conversion. A functionality to search for special postures of interest and display the frames in an overview makes it easy to analyze difference phenomenas in Conversation Analysis.

Authors Index