LREC 2014
Reykjavik

NINTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION

Held under the Patronage of UNESCO, the United Nations Educational, Scientific and Cultural Organization

MAY 26 – 31, 2014

HARPA CONFERENCE CENTER
REYKJAVIK, ICELAND

WORKSHOP ABSTRACTS

Editors: Please refer to each single workshop list of editors.

Assistant Editors: Sara Goggi, Hélène Mazo

LREC 2014, NINTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION

Title: LREC 2014 Workshop Abstracts

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

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

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

ISBN 978-2-9517408-8-4
EAN 9782951740884

TABLE OF CONTENTS

5th International Workshop on EMOTION, SOCIAL SIGNALS, SENTIMENT & LINKED OPEN DATA

26-27 May 2014

ABSTRACTS

Editors:

Björn Schuller, Paul Buitelaar, Laurence Devillers, Catherine Pelachaud, Thierry Declerck, Anton Batliner, Paolo Rosso, Seán Gaines

Workshop Programme

26 May 2014

14:00 – 15:00 **Keynote I** (Chair: Björn Schuller)

Walter Daelemans (University of Antwerp, The Netherlands), *Profiling and sentiment mining for detecting threatening situations in social networks: the AMiCA project*

15:00 – 16:00 **Session 1: Markup and Linked Data** (Chair: Walter Daelemans)

Felix Burkhardt, Christian Becker-Asano, Edmon Begoli, Roddy Cowie, Gerhard Fobe, Patrick Gebhard, Abe Kazemzadeh, Ingmar Steiner and Tim Llewellyn, *Application of EmotionML*

Gabriela Vulcu, Paul Buitelaar, Sapna Negi, Bianca Pereira, Mihael Arcan, Barry Coughlan, Fernando J. Sanchez and Carlos A. Iglesias, *Generating Linked-Data based Domain-Specific Sentiment Lexicons from Legacy Language and Semantic Resources*

16:00 – 16:30 Coffee break

16:30 – 18:00 **Session 2: Spoken Language** (Chair: Laurence Devillers)

Anna Prokofieva and Julia Hirschberg, *Hedging and Speaker Commitment*

Björn Schuller, Yue Zhang, Florian Eyben and Felix Weninger, *Intelligent Systems' Holistic Evolving Analysis of Real-life Universal Speaker Characteristics*

Zixing Zhang, Florian Eyben, Jun Deng and Björn Schuller, *An Agreement and Sparseness-based Learning Instance Selection and its Application to Subjective Speech Phenomena*

27 May 2014

09:00 – 10:30 **Keynote II and Plenary Discussion** (Chair: Paul Buitelaar)

Carlos Iglesias (Universidad Politécnica de Madrid, Spain), *A linked data approach for describing sentiments and emotions*

Plenary Discussion: *W3C Community Group on Linked Data Models for Emotion and Sentiment Analysis*

10:30 – 11:00 Coffee break

11:00 – 13:00 **Session 3: Corpora and Data Collection** (Chair: Thierry Declerck)

Véronique Aubergé, Yuko Sasa, Nicolas Bonnefond, Brigitte Meillon, Tim Robert, Jonathan Rey-Gorrez, Adrien Schwartz, Leandra Antunes, Gilles De Biasi, Sybille Caffiau and Florian Nebout, *The EEE corpus: socio-affective "glue" cues in elderly-robot interactions in a Smart Home with the EmOz platform*

Mohamed A. Sehili, Fan Yang, Violaine Leynaert and Laurence Devillers, *A corpus of social interaction between NAO and elderly people*

Kateřina Veselovská, *Fear and Trembling: Annotating Emotions in Czech Holocaust Testimonies*

Heather Pon-Barry, *Using Ambiguous Handwritten Digits to Induce Uncertainty*

13:00 – 14:00 Lunch break

14:00 – 16:00 **Session 4: Social Networks** (Chair: Carlos Iglesias)

Eshrag Refaee and Verena Rieser, *Can We Read Emotions from a Smiley Face? Emoticon-based Distant Supervision for Subjectivity and Sentiment Analysis of Arabic Twitter Feeds*

Cristina Bosco, Leonardo Allisio, Valeria Mussa, Viviana Patti, Giancarlo Ruffo, Manuela Sanguinetti and Emilio Sulis, *Detecting Happiness in Italian Tweets: Towards an Evaluation Dataset for Sentiment Analysis in Felicittà*

Erik Tjong Kim Sang, *Using Tweets for Assigning Sentiments to Regions*

Francisco Rangel, Irazú Hernández, Paolo Rosso and Antonio Reyes, *Emotions and Irony per Gender in Facebook*

16:00 – 16:30 Coffee break

16:30 – 18:00 **Session 5: Written Language** (Chair: Chloé Clavel)

Ekaterina Volkova and Betty J. Mohler, *On-line Annotation System and New Corpora for Fine-Grained Sentiment Analysis of Text*

Elizabeth Baran, *Correlating Document Sentiment Scores with Web-Sourced Emotional Response Polls for a More Realistic Measure of Sentiment Performance*

Caroline Langlet and Chloé Clavel, *Modelling user's attitudinal reactions to the agent utterances: focus on the verbal content*

Workshop Organisers

Björn Schuller	Imperial College London, UK
Paul Buitelaar	NUI Galway, Ireland
Laurence Devillers	U. Sorbonne/CNRS-LIMSI, France
Catherine Pelachaud	CNRS-LTCI, France
Thierry Declerck	DFKI, Germany
Anton Batliner	FAU/TUM, Germany
Paolo Rosso	PRHLT, U. Politèc. Valencia, Spain
Seán Gaines	Vicomtech-IK4, Spain

Workshop Programme Committee

Rodrigo Agerri	EHU, Spain
Noam Amir	Tel-Aviv U., Isreal
Elisabeth André	U. Augsburg, Germany
Alexandra Balahur-Dobrescu	ISPRA, Italy
Cristina Bosco	U. Torino, Italy
Felix Burkhardt	Deutsche Telekom, Germany
Carlos Busso	UT Dallas, USA
Rafael Calvo	U. Sydney, Australia
Erik Cambria	NUS, Singapore
Antonio Camurri	U. Genova, Italy
Mohamed Chetouani	UPMC, France
Montse Cuadros	VicomTech, Spain
Francesco Danza	Expert System, Italy
Thierry Dutoit	U. Mons, Belgium
Julien Epps	NICTA, Australia
Anna Esposito	IIASS, Italy
Francesca Frontini	CNR, Italy
Hatice Gunes	Queen Mary U., UK
Hayley Hung	TU Delft, the Netherlands
Carlos Iglesias	UPM, Spain
Isa Maks	VU, the Netherlands
Daniel Molina	Paradigma Tecnologico, Spain
Monica Monachini	CNR, Italy
Shrikanth Narayanan	USC, USA
Viviana Patti	U. Torino, Italy
German Rigau	EHU, Spain
Fabien Ringeval	U. Fribourg, Switzerland
Massimo Romanelli	Attensity EUROPE, Germany
Albert Ali Salah	Boğaziçi University, Turkey
Metin Sezgin	Koc U., Turkey
Carlo Strapparava	FBK, Italy
Jianhua Tao	CAS, P.R. China
Tony Veale	UCD, Ireland
Michel Valstar	U. Nottingham, UK
Alessandro Vinciarelli	U. Glasgow, UK
Piek Vossen	VU, the Netherlands

Emotion, Social Signals, Sentiment & Linked Open Data: A Short Introduction

The fifth instalment of the highly successful series of Corpora for Research on Emotion held at the last LRECs (2006, 2008, 2010, 2012) aims to help further bridging the gap between research on human emotion, social signals and sentiment from speech, text, and further modalities, and low availability of language and multimodal resources and labelled data for learning and testing.

As usually rather labels than the actual data are sparse, this year emphasis was put also on efficient community-shared and computer-supported labelling approaches and on cross-corpora experiments. Following LREC 2014's hot topics of Big Data and Linked Open Data in particular also methods for semi-automated and collaborative labelling of large data archives such as by efficient combinations of active learning and crowd sourcing are featured in this edition – in particular also for combined annotations of emotion, social signals, and sentiment. Multi- and cross-corpus studies (transfer learning, standardisation, corpus quality assessment, etc.) were further considered as highly relevant, given their importance in order to test the generalisation power of models.

A further main motivation for this year's workshop was to survey and promote the uptake of Linked Data in emotion, sentiment & social signal analysis research and applications. Linked Open Data is an increasingly wide-spread methodology for the publishing, sharing and interlinking of data sets. In the context of this workshop we were also interested in reports on and experiences with the use of Linked Open Data in the context of emotion, social signals, and sentiment in analysis projects and applications.

As before, also the multimodal community was invited and encouraged to contribute new corpora, perspectives and findings – emotion, sentiment, and social behaviour are multimodal and complex and there is still an urgent need for sufficient naturalistic uni- and multimodal data in different languages and from different cultures.

From the papers received, 16 were selected for the final programme (rejecting six) by the 36 members of the technical programme committee and the eight organisers. The accepted contributions were all selected as oral presentation and come from a total of 65 authors. They were grouped into the five groups *markup (languages) and linked data* (two papers), *spoken language* (three papers), *corpora and data collection* (four papers), *social networks* (four papers), and *written language* (three papers). Obviously, several of the papers fall under multiple of these headings and other groupings could have been thought off.

From the 16 accepted contributions one was selected as best paper by the technical program committee and organisers based on the review results and a rigorous second screening – contributions including members of the organising committee were not eligible for fairness reasons. This best paper award was given to Véronique Aubergé, Yuko Sasa, Nicolas Bonnefond, Brigitte Meillon, Tim Robert, Jonathan Rey-Gorrez, Adrien Schwartz, Leandra Antunes, Gilles De Biasi, Sybille Caffiau and Florian Nebout for their outstanding and inspiring introduction and efforts of and around *The EEE corpus: socio-affective "glue" cues in elderly-robot interactions in a Smart Home with the EmOz platform*.

Two keynote speeches by distinguished researchers crossing the communities further focused on the above named topics of particular interest: Walter Daelemans's (University of Antwerp, The Netherlands) talk *Profiling and sentiment mining for detecting threatening situations in social*

networks: the AMiCA project introduced findings from a larger project. The second speech given by Carlos Iglesias (Universidad Politécnica de Madrid, Spain) was entitled *A linked data approach for describing sentiments and emotions*, and followed by a plenary discussion around the *W3C Community Group on Linked Data Models for Emotion and Sentiment Analysis*.

The organisers are further grateful for the sponsorship of the Association for the Advancement of Affective Computing (AAAC, former HUMAINE Association) and the SSPNet. The workshop was further partially organised in the context of and received funding from the following European projects: ASC-Inclusion (http://www.asc-inclusion.eu), EuroSentiment (http://eurosentiment.eu), iHEARu (http://www.ihearu.eu), ilhaire (http://www.ilhaire.eu/), LIDER (http://lider-project.eu/), OpeNER (http://www.opener-project.org), TARDIS (http://www.tardis-project.eu), TrendMiner (http://www.trendminer-project.eu), and WiQ-Ei. The responsibility lies with the organisers and authors.

To conclude, we would like to thank all the dedicated members of the technical program committee, the sponsors, ELRA, and of course all authors for an inspiring and exciting workshop and proceedings.

Björn Schuller, Paul Buitelaar, Laurence Devillers, Catherine Pelachaud, Thierry Declerck, Anton Batliner, Paolo Rosso, Seán Gaines

Organisers of ES³LOD 2014

Keynote I

Monday 26 May, 14:00 – 15:00
Chairperson: Björn Schuller

Profiling and sentiment mining for detecting threatening situations in social networks: the AMiCA project

Walter Daelemans (CLiPS, University of Antwerp)

Session 1: Markup and Linked Data

Monday 26 May, 15:00 – 16:00
Chairperson: Björn Schuller

Application of EmotionML

Felix Burkhardt, Christian Becker-Asano, Edmon Begoli, Roddy Cowie, Gerhard Fobe, Patrick Gebhard, Abe Kazemzadeh, Ingmar Steiner and Tim Llewellyn

Abstract
We present EmotionML1, a new W3C recommendation to represent emotion related states in data processing systems as well as a series of concrete implementations that utilize EmotionML. EmotionML was developed by a subgroup of the W3C MMI (Multimodal Interaction) Working Group chaired by Deborah Dahl in a first version from approximately 2005 until 2013, most of this time the development was lead by Marc Schröder. The first part of this paper deals with a short summary of EmotionML by describing selected aspects and the procedure and thinking behind its development. The second half introduces a number of applications that integrated EmotionML and were submitted as implementation reports during the W3C recommendation track process.

Generating Linked-Data based Domain-Specific Sentiment Lexicons from Legacy Language and Semantic Resources

Gabriela Vulcu, Paul Buitelaar, Sapna Negi, Bianca Pereira, Mihael Arcan, Barry Coughlan, Fernando J. Sanchez and Carlos A. Iglesias

Abstract
We present a methodology for legacy language resource adaptation that generates domain-specific sentiment lexicons organized around domain entities described with lexical information and sentiment words described in the context of these entities. We explain the steps of the methodology and we give a working example of our initial results. The resulting lexicons are modelled as Linked Data resources by use of established formats for Linguistic Linked Data (lemon, NIF) and for linked sentiment expressions (Marl), thereby contributing and linking to existing Language Resources in the Linguistic Linked Open Data cloud.

Session 2: Spoken Language

Monday 26 May, 16:30 – 18:00
Chairperson: Laurence Devillers

Hedging and Speaker Commitment

Anna Prokofieva and Julia Hirschberg

Abstract
Hedging is a phenomenon in which a speaker communicates a lack of commitment to what they are saying. Hedges occur quite commonly in text and speech and have been associated with many discourse actions, such as trying to save face, avoid criticism and show politeness. Currently, there is only one available corpus annotated for hedges - the BioScope corpus of articles and abstracts in the biomedical field created for the CONLL 2010 Shared Task. Since hedging is an important conversational phenomenon, we believe it is necessary to create a resource for investigating hedges in speech. To this end, we have expanded upon the CONLL 2010 guidelines and propose a semi-automated scheme for annotating several speech corpora for this phenomenon. We hope that this effort allows further investigation into the correlations between hedging and various social aspects of dialogue.

Intelligent Systems' Holistic Evolving Analysis of Real-life Universal Speaker Characteristics

Björn Schuller, Yue Zhang, Florian Eyben and Felix Weninger

Abstract
In this position paper we introduce the FP7 ERC starting grant project “Intelligent systems’ Holistic Evolving Analysis of Real-life Universal speaker characteristics” – iHEARu. Further, a first approach of multilabel classification in parallel tasks is proposed to analyse speech in a holistic fashion, learning how speaker characteristics influence each other. It is shown in this work that learning these labels and tasks simultaneously and jointly outperforms the common approach of learning them separately.

An Agreement and Sparseness-based Learning Instance Selection and its Application to Subjective Speech Phenomena

Zixing Zhang, Florian Eyben, Jun Deng and Björn Schuller

Abstract
Like in other pattern recognition tasks, redundant instances in subjective speech phenomena may cause increased training time and performance degradation of a classifier. Instance selection, aiming to discard some 'troublesome' instances and choose the most informative ones, is a way to solve this issue. We thus propose a tailored algorithm based on human Agreement levels of labelling and class Sparseness for learning Instance Selection – ASIS for short. Extensive experiments on a standard speech emotion recognition task show the effectiveness of ASIS, indicating that by selecting only 30 % of the training set, the system performance significantly outperforms training on the whole training set without instance balancing. In terms of performance it remains comparable to the classifier trained with instance balancing, but at a fraction of the training material.

Keynote II and Plenary Discussion

Tuesday 27 May, 9:00 – 10:30
Chairperson: Paul Buitelaar

A linked data approach for describing sentiments and emotions

Carlos A. Iglesias (Universidad Politécnica de Madrid),

W3C Community Group on Linked Data Models for Emotion and Sentiment Analysis

Plenary Discussion lead by: Paul Buitelaar, Carlos A. Iglesias, Björn Schuller

Session 3: Corpora and Data Collection

Tuesday 27 May, 11:00 – 13:00
Chairperson: Thierry Declerck

The EEE corpus: socio-affective "glue" cues in elderly-robot interactions in a Smart Home with the EmOz platform

Véronique Aubergé, Yuko SASA, Nicolas Bonnefond, Brigitte Meillon, Tim Robert, Jonathan Rey-Gorrez, Adrien Schwartz, Leandra Antunes, Gilles De Biasi, Sybille Caffiau and Florian Nebout

Abstract
The aim of this study is to give a glance at interactions in a Smart Home prototype between the elderly and a companion robot that is having some socio-affective language primitives as the only vector of communication. Through a Wizard of Oz platform (EmOz), a robot is introduced as an intermediary between the technological environment and some elderly who have to give vocal commands to the robot to control the Smart Home. The robot vocal productions increases progressively by adding prosodic levels: (1) no speech, (2) pure prosodic mouth noises supposed to be the "glue's" tools, (3) lexicons with supposed "glue" prosody and (4) subject's commands imitations with supposed "glue" prosody. The elderly subjects' speech behaviors confirm the hypothesis that the socio-affective "glue" effect increase towards the prosodic levels, especially for socio-isolated people. This paper will specifically focus on the experiment script and the Wizard of Oz tool developed to observe the spontaneous human-robot interactions.

A corpus of social interaction between NAO and elderly people

Mohamed A. Sehili, Fan Yang, Violaine Leynaert and Laurence Devillers

Abstract
This paper presents a corpus featuring social interaction between elderly people in a retirement home and the humanoid robot Nao. This data collection is part of the ROMEO2 project which is following the French ROMEO project (http://projetromeo.com//index_en.html) [Delaborde & Devillers, 2010, Buendia & Devillers, 2013]. The goal of the project is to develop a humanoid robot that can act as a comprehensive assistant for persons suffering from loss of autonomy. In this perspective, the robot is able to assist a person in their daily tasks when they are alone. The aim of this study is to design an affective interactive system driven by interactional, emotional and personality markers. Our paper will present the data collection protocol and scenarios, the corpus collected (27 people, average age: 85) and the results on analyses of questionnaires (Satisfaction, OCEAN) and some annotations of the commitment level (laughers, smiles, etc).

Fear and Trembling: Annotating Emotions in Czech Holocaust Testimonies

Kateřina Veselovská

Abstract
In this paper, we introduce the Visual History Archive of the USC Shoah Foundation as a multimodal data resource for sentiment analysis in Czech, but potentially all thirty three languages it contains. We take the opportunity of having both physical access to these unique data and the well-established research group on sentiment analysis at Charles University in Prague. Our aim is to provide methodology for sentiment annotation of these multimodal data combining subjectivity detection within a treebank with spoken term detection.

Using Ambiguous Handwritten Digits to Induce Uncertainty

Heather Pon-Barry

Abstract
The lack of ground truth labels is a significant challenge in the field of automatic recognition of emotion and affect. The most common approach to acquiring affect labels is to ask a panel of listeners to rate a corpus of spoken utterances along one or more dimensions of interest. In this paper, we describe a method that uses ambiguous handwritten digits for the purpose of inducing natural uncertainty. Using a crowdsourcing approach, we quantify the legibility of each handwritten digit. These images are integrated into visual stimuli that will be used in a question-answering lab experiment for eliciting spontaneous spoken answers of varying levels of certainty. While we cannot control a speaker's actual internal level of certainty, the use of these stimuli provides an approximation.

Session 4: Social Networks

Tuesday 27 May, 14:00 – 16:00
Chairperson: Carlos Iglesias

Can We Read Emotions from a Smiley Face? Emoticon-based Distant Supervision for Subjectivity and Sentiment Analysis of Arabic Twitter Feeds

Eshrag Refaee and Verena Rieser

Abstract
The growth of social media, especially as a source for analysis, has resulted in a two-fold challenge: managing the costs processing all of that data, as well as developing new ways to make sense of it. And, of course, in the small world in which we live, one needs to be able to handle multiple languages and idioms equally well. In this work we explore different approaches to Subjectivity and Sentiment Analysis (SSA) of Arabic tweets. SSA aims to determine the attitude of the speaker with respect to some topic, e.g. objective or subjective, or the overall contextual polarity of an utterance, e.g. positive or negative. Compared to other languages, such as English, annotated data available for research on Arabic SSA is sparse. Creating a new data set is costly, and, as we will show in the following, learning from small data sets does not cover the wide scope of topics discussed on twitter. This research is the first to explore distant supervision approaches for automatic SSA classification for Arabic social networks.

Detecting Happiness in Italian Tweets: Towards an Evaluation Dataset for Sentiment Analysis in Felicittà

Cristina Bosco, Leonardo Allisio, Valeria Mussa, Viviana Patti, Giancarlo Ruffo, Manuela Sanguinetti and Emilio Sulis

Abstract
This paper focuses on the development of a gold standard corpus for the validation of Felicitt`a, an online platform which uses Twitter as data source in order to estimate and interactively display the degree of happiness in the Italian cities. The ultimate goal is the creation of an Italian reference Twitter dataset for sentiment analysis that can be used in several frameworks aimed at detecting sentiment from big data sources. We will provide an overview of the reference corpus created for evaluating Felicitt`a, with a special focus on the issues raised from its development, from inter-annotator agreement analysis and on implications for the further development of the corpus.

Using Tweets for Assigning Sentiments to Regions

Erik Tjong Kim Sang

Abstract
We derive a sentiment lexicon for Dutch tweets and apply the lexicon for classifying Dutch tweets as positive, negative or neutral. The classifier enables us to test what regions in the Netherlands and Flanders express more positive sentiment on Twitter than others. The results reveal sentiment differences between Flemish and Dutch provinces, and expose municipalities which are a lot more negative than their neighborhood. The results of this study can be used for finding areas with local issues that might be expressed in tweets.

Emotions and Irony per Gender in Facebook

Francisco Rangel, Irazú Hernández, Paolo Rosso and Antonio Reyes

Abstract
Our habits are changing, we are no longer customers searching for products but users looking for new experiences. Social Media even accentuate such changes. The emotional aspect of the life is acquiring a growing importance. Thus, the need of affective processing acquires a new dimension nowadays in order to know what users want and need. We are interested in social media since we are interested in everyday language and how it reflects basic social, emotional and personal processes. Furthermore, in social media users reflect what they want and need without restrictions and liberty of expression. But there is a lack of annotated resources on affectivity when we talk about social media texts. Even more if we focus on Spanish language, no matter its good penetration in Internet. We focused on Facebook as representative of social media because it is massively used by people, where they express their thoughts freely and without editorial guidelines unlike traditional media like newsletters and with spontaneity unlike blogs. Thus the expected affectivity in such media is very high. Facebook also allows us to obtain demographics such as gender, unlike similar media like Twitter. The paper describes a dataset collected from Facebook, in Spanish and labelled with emotions, irony and gender of authors. We describe we describe the corpus, how the data was collected and annotated, and the inter-annotator agreement. We analyse the corpus and we present statistics about emotions and irony per gender. We describe how the corpus will be distributed for research purposes. Finally we draw some conclusions.

Session 5: Written Language

Tuesday 27 May, 16:30 – 18:00
Chairperson: Chloé Claval

On-line Annotation System and New Corpora for Fine-Grained Sentiment Analysis of Text

Ekaterina Volkova and Betty J. Mohler

Abstract
We present a new online tool designed for fine-grained sentiment annotation that could be useful to the linguistic community for fast and effortless collection of high-quality annotations. The tool is to be available for both, researchers and annotators, is web-based and does not require any specific software. It allows the annotators to encode their emotional perception of text with the help of a rich set of emotions, indicate the intensity for each emotion instance and mark the word that bears most of the emotional charge. We demonstrate the work of the annotation system by collecting two corpora of annotation texts: a large German corpus annotated by two people, and a smaller corpus of English fairy tales with seven annotators for each story. We show that the consensus method of establishing a gold-standard annotation results in an annotation rich with emotion labels. The annotation tool as well as both corpora are available upon request.

Correlating Document Sentiment Scores with Web-Sourced Emotional Response Polls for a More Realistic Measure of Sentiment Performance

Elizabeth Baran

Abstract
Sentiment analysis as a multi-faceted problem that we often try to reduce down to measures of correct and incorrect. We propose instead a method for evaluating document-level sentiment polarity that uses correlation statistics. We test our document-level sentiment detection engine, which uses a sentiment phrase dictionary, against web-sourced emotional response poll data in two languages, Italian and Chinese. We look at how our engine's document-level sentiment scores correlate with the sentiment poll data. We then show how these correlations change with additions or modifications to our sentiment phrase dictionary.

Modelling user's attitudinal reactions to the agent utterances: focus on the verbal content

Caroline Langlet and Chloé Clavel

Abstract
With the view to develop a module for the detection of user's sentiments in a human-agent interaction, the present paper proposes to go beyond the classical positive vs. negative distinction used in sentiment analysis and provides a model of user's attitudes in verbal content. This model, grouned on the Martin and White's Appraisal Theory, deal with the user's attitude in the interaction. We present here the key features of this model and we analyse the annotations obtained by confronting it with the SEMAINE corpus.

Automatic and Manual Metrics for Operational Translation Evaluation

26 May 2014

ABSTRACTS

Editors:

Keith J. Miller, Lucia Specia, Kim Harris, Stacey Bailey

Workshop Programme

08:45 – 09:30 Welcome and Introduction by Workshop Organizers

09:30 – 10:30 Talks Session 1

Joke Daems, Lieve Macken and Sonia Vandepitte, *Two Sides of the Same Coin: Assessing Translation Quality in Two Steps Through Adequacy and Acceptability Error Analysis*

Leonid Glazychev, *How to Reliably Measure Something That's Not Completely Objective: A Clear, Working and Universal Approach to Measuring Language Quality*

Mihaela Vela, Anne-Kathrin Schumann and Andrea Wurm, *Translation Evaluation and Coverage by Automatic Scores*

Arle Lommel, Maja Popović and Aljoscha Burchardt, *Assessing Inter-Annotator Agreement for Translation Error Annotation*

10:30 – 11:00 Coffee break

11:00 – 13:00 Talks Session 2

Marianne Starlander, *TURKOISE: A Mechanical Turk-based Tailor-made Metric for Spoken Language Translation Systems in the Medical Domain*

Caitlin Christianson, Bonnie Dorr and Joseph Olive, *MADCAT Evaluation Approach: Operational Accuracy of MT applied to OCR*

Ekaterina Stambolieva, *Continuous Operational Evaluation of Evolving Proprietary MT Solution's Translation Adequacy*

Lars Ahrenberg, *Chunk Accuracy: A Simple, Flexible Metric for Translation Quality*

Michael Carl and Moritz Schaeffer, *Word Transition Entropy as an Indicator for Expected Machine Translation Quality*

Douglas Jones, Paul Gatewood, Martha Herzog and Tamas Marius, *A New Multiple Choice Comprehension Test for MT* and *Standardized ILR-Based* and *Task-Based Speech-to-Speech MT Evaluation*

Lena Marg, *Rating Evaluation Methods through Correlation*

Federico Gaspari, Antonio Toral, Arle Lommel, Stephen Doherty, Josef van Genabith and Andy Way, *Relating Translation Quality Barriers to Source-Text Properties*

13:00 – 14:00 Lunch break

14:00 – 15:00 Hands-On Session 1

15:00 – 16:00 Hands-On Session 2

16:00 – 16:30 Coffee break

16:30 – 17:30 Hands-On Session 3

17:30 – 18:00 Discussion, Potential for Future Collaboration, Next Steps, and Conclusion

Organizing Committee and Editors

Keith J. Miller	The MITRE Corporation
Lucia Specia	University of Sheffield
Kim Harris	text&form GmbH, Germany
Stacey Bailey	The MITRE Corporation

Introduction Session

08:45 – 09:30
Chairperson: Lucia Specia

Understanding Stakeholder Requirements for Determining Translation Quality

Tyler Snow and Alan Melby
E-mail: tylerasnow@gmail.com, alan.melby@gmail.com

This paper presents the results of a large-scale study on the translation-related language quality assessment practices of language service providers, content creators who purchase translation, and free-lance translators. Conducted by the Globalization and Localization Association (GALA) as part of the EU-funded QTLaunchPad Project, this study is intended to provide concrete feedback to influence the development of the Multidimensional Quality Metrics (MQM) system for analyzing translation quality. By specifying what the "real-world" requirements for a translation quality assessment system are, it will help ensure that MQM is aligned with industry best practice and is flexible enough to meet the requirements of the full range of potential users in industry and research.

The study began with a survey sent out to thousands of individuals in the above-mentioned stakeholder segments around the world concerning quality management as applied to their translation activities. Approximately 300 persons participated in the survey, and approximately 60 percent of those indicated they would be interested in follow-up interviews. Key findings include:

(1) There is no industry consensus on appropriate quality processes, and assessment processes are highly diverse, ranging from informal, subjective readings to highly rigorous, analytic approaches. There are currently no widely accepted best practices.

(2) The most common method involves "spot checks" conducted on small samples of translated data to determine whether texts are "good enough" or need additional remediation.

(3) Most of those surveyed use "analytic" quality assessment methods that evaluate the translated text closely to identify and quantify specific errors in the text. Less common alternatives include a "holistic" approach that involves rating the overall translation on one or more dimensions.

(4) The most common specific metrics today are either in-house adaptations of the LISA QA Model or ones built into tools such as CAT tools or purpose-built translation quality-checking tools.

(5) Many quality assessment processes use a scorecard to aid in evaluation. Evaluators go through the text to mark and categorize errors, information about which is entered into the scorecard to calculate a quality score (usually expressed as a percentage value).

(6) The most frequently checked issues are: technical issues related to internationalization/ localization engineering, accuracy (e.g., mistranslation), fluency (e.g., linguistic features and grammar), terminology compliance, typography, compliance with legal requirements, and consistency. But many stakeholders wish that metrics would address other features such as offensiveness, readability, functionality of code (to ensure that localization has not "broken" it), productivity, and adherence to specifications.

Understanding these quality processes and the requirements that various stakeholder groups have within the translation process is crucial for improving the quality assessment of translation and providing results that accurate reflect the "quality" of texts in real-world situations. This presentation provides an overview of the findings of the survey and qualitative interviews, specifically as they relate to the MQM system for defining quality metrics. It will identify the most common error categories found and discuss how they are used in industry settings and the practical issues that the various stakeholder segments experience in their efforts to define, determine, and assure quality.

Automated and Task-Based Evaluation of the Effects of Machine Translation Domain Tuning on MT Quality, Post-editing, and Human Translation

Stacey Bailey and Keith J. Miller
E-mail: sbailey@mitre.org, keith@mitre.org

Domain tuning (DT) is the process of tailoring a machine translation (MT) system to better handle data relating to a particular topic area, either by training the MT system with data that is representative of the topic's subject matter (e.g., scientific and technical literature) or by adding terminology that is relevant to that subject matter. While DT can improve the quality of MT output, knowing how, when, and to what extent users should invest in developing corpus and lexical resources for DT is unclear. This research begins to address these questions by investigating the effects of domain-tuning on the quality of the output of two commercial MT systems.

This research evaluates two approaches to machine translation domain tuning (MTDT): (1) training a custom engine using parallel, domain data and (2) lightweight tuning using domain-specific glossaries. The research combined automatic evaluation and in-depth task-based evaluation of Chinese-to-English translation in the cyber domain. This study provided a 3-way comparison between 1) post-editing MT output from two commercial MT systems, 2) human translation of texts with no MT, and 3) human translation without MT but with domain term translations provided.

The three working hypotheses were that 1) DT improves the quality of machine translation output over baseline capabilities, as measured by automatic evaluation metrics, 2) Human translation time can be reduced by requiring human translators to post-edit the output of domain-tuned MT systems and 3) The linguistic quality of the target language document can be improved by requiring human translators to post-edit the output of domain-tuned MT systems as opposed to starting with source text only. It was hypothesized the post-editing DT would improve speed and quality of translation as compared to both post-editing of baseline MT and human translation without MT.

For each MT engine, there were four engine variations compared, yielding a total of eight MT test conditions: Post-editing using (1) the MT engine without any DT. (2) the MT engine plus lightweight DT with a found domain-specific lexicon. (3) the MT engine plus a statistically retrained engine based on the training data, and (4) the MT engine plus both a statistically retrained engine and a found lexicon. There were two additional conditions compared to these MT conditions: (5) Manual translation that does not use MT but does use a domain-specific lexicon for highlighting found terms with glosses provided for the translator. (6) Manual translation with no MT or term highlighting.

16 participants were given abstracts that included just the source Chinese text or the source text plus either the output of the MT (one of the MT test conditions) or the manually highlighted terms. They were asked to correct the MT output or produce a final translation from scratch. Translation times were recorded, and after each translation, the participants were given a survey about the utility of the resources provided and their opinions of the translation quality.

The results suggest that, generally speaking, DT can improve performance on automatic MT metrics, but it is not straightforward to predict whether a particular type of DT will definitely improve performance on a given domain. For some conditions and metrics, the performance dropped with DT. With respect to translation rates, results were also mixed. Rates were faster for some MT conditions and slower for others. Most notably, it was slowest on output from the two MT conditions based on the most involved DT.

Finally, six quality control (QC) translators were given the Chinese segments with the collected English translations. The QC-ers reviewed the source segment, rated each translation, and counted errors. Follow-on work will correlate these data with the automatic metrics and time data.

Talks Session 1

09:30 – 10:30
Chairperson: Lucia Specia

Two Sides of the Same Coin: Assessing Translation Quality in Two Steps through Adequacy and Acceptability Error Analysis

Joke Daems, Lieve Macken, Sonia Vandepitte
E-mail: joke.daems@ugent.be, lieve.macken@ugent.be, sonia.vandepitte@ugent.be

A translator has to find the balance between adhering to the norms of the source text (adequacy) and respecting the norms of the target text (acceptability) (Toury, 1995). The quality of a translation can then be judged on its (non-)adherence to these norms. This is a common quality judgment for machine translation, where evaluators give translated segments an adequacy and acceptability (sometimes 'fluency') score on a scale from one to five (White, 1995).

When looking at translation quality assessment through error analysis, however, the dichotomy between acceptability and adequacy is not always as distinct. Existing metrics do provide error categories relating to both types of issues. For example, QTLaunchPad's MQM has a category for fluency and one for accuracy; TAUS suggests a category for accuracy, one for terminology, and two categories that could relate to acceptability (language and style); FEMTI proposes suitability, accuracy and wellformedness; and MeLLANGE offers categories for language and content transfer. Yet these categories are all part of one and the same evaluation step: evaluators have to identify issues and assign the correct category to these issues. Research has shown that deciding whether an error belongs to adequacy or acceptability is one of the most difficult aspects of error analysis for human annotators, together with having to assign an error weight to each error instance (Stymne & Ahrenberg, 2012).

We therefore propose facilitating the error annotation task by introducing an annotation process which consists of two separate steps that are similar to the ones required in the European Standard for translation companies EN 15038: an error analysis for errors relating to acceptability (where the target text as a whole is taken into account, as well as the target text in context), and one for errors relating to adequacy (where source segments are compared to target segments). We present a fine-grained error taxonomy suitable for a diagnostic and comparative analysis of machine translated-texts, post-edited texts and human translations. Categories missing in existing metrics have been added, such as lexical issues, coherence issues, and text type-specific issues. Annotator subjectivity is reduced by assigning error weights to each error category beforehand, which can be tailored to suit different evaluation goals, and by introducing a consolidation step, where annotators discuss each other's annotations.

The approach has been tested during two pilot studies with student translators who both post-edited and translated different texts. Inter-annotator agreement shows that the proposed categorization is clear and that it is necessary to include a consolidation phase. Annotations after consolidation were used to analyze the most common errors for each method of translation and to provide an average error score per word for each text. In a next phase, the annotations were manually grouped into source text-related error sets: a source text passage and the translations for that passage that contain errors. Error sets allow for a diagnostic evaluation: which source text-segments that were problematic for machine translation are still problematic after post-editing and how? How many and which post-editing errors originate from the machine translation output?

Though the approach in its current form requires much time and human effort (the annotation process in itself costs around 45 minutes for 150 words of a new MT text, with acceptability annotations requiring the most time: 30 minutes), it does provide rich data needed to improve translation quality. Familiarity with a text can seriously decrease annotation time, and the time for HT or PE if also lower than for MT. We are currently optimizing the annotation process to increase the speed and reduce manual effort, and we believe that the processing of the annotations and the creation of the error sets can, at least in part, be automated.

How to Reliably Measure Something That's Not Completely Objective:
A Clear, Working and Universal Approach to Measuring Language Quality

Leonid Glazychev
E-mail: leonidg@logrus.net, lglazychev@outlook.com

While everybody needs to measure language quality, and numerous practical models as well as theoretical approaches have been developed over decades, all these models and approaches were concentrating on particular factors to measure and their relative weights, i.e. what is important and what is not. At the same time practical, real-world solutions targeted at human reviewers and applicable beyond the MT domain, when we have to deal with new translations of unknown origin (either human or [post-edited] MT) with no reference translations available, are scarce. Creating one requires providing answers to the following questions:

- How exactly can we reliably measure something that is not completely objective by design?
- How trustworthy the results of each particular human review are, and what is the best way to analyse and interpret them?
- How to develop the quality measurement approach/metric that is not simply justified, flexible and reliable enough, but can also be utilized in real life as part of the production process?

The presentation outlines the approach developed by the author at Logrus International Corporation and based on years of research and practical work on clients' projects. The suggested model is flexible (can be easily adapted to particular type of content or requirements), universal (can be applied to both human and machine translation) and practical (can be implemented as part of the production process).

The concept is based on selecting primary factors influencing the perception and priorities of the target audience, separating global and local issues and dividing all quality-related factors into three basic categories: objective, semi-objective and subjective. Each category is described in detail, including its nature, limitations and specifics. This classification is paired with a multidimensional approach to quality built around four "cornerstones": Adequacy, Readability, Technical Quality, and Major Errors.

The presentation provides concrete recommendations on the process, which includes:

- Applying threshold-based (pass/fail) criteria for most important global, semi-objective factors, such as content adequacy to the original and overall readability (fluency).
- Considering major errors (grossly distorting the meaning, creative offensive statements, etc.)
- Counting and classifying technical errors in materials that passed the first two tests and applying error-weighting templates appropriate for the occasion.

The presentation gives a deeper insight into interpreting quality review results and explains why this hybrid approach combining threshold-based (rubric) and regular quantitative criteria is optimal, discusses grading scales, etc. Practical details include the following:

- Why one should not over-rely on particular figures
- Why it is not recommended to wrap everything into a single quality evaluation grade, and why we need to substitute it with four different values defining the "quality square".
- Why is the scale used for grading so important

Both the "full" and "light" quality evaluation models are presented. The latter is significantly less effort-consuming and consequently less precise, but is invaluable for public initiatives involving unpaid community-sourced effort or for cases of quality evaluation on a shoestring budget.

All recommendations are illustrated using actual statistical results obtained through a community review of the localized version of a popular website by 18 professional translators.

Human Translation Evaluation and its Coverage by Automatic Scores

Mihaela Vela, Anne-Kathrin Schumann, Andrea Wurm
E-mail: m.vela@mx.uni-saarland.de, anne.schumann@mx.uni-saarland.de, a.wurm@mx.uni-saarland.de

Approaches to the evaluation of machine translation output are numerous and range from fully automatic quality scoring to efforts aimed at the development of "human" evaluation scores. The goals for which such evaluations are performed are manifold, covering system optimisation and benchmarking as well as the integration of MT engines into industrially deployable translation workflows. The discipline of translation studies, on the other hand, can look back onto a long line of thought on the quality of translations. While the discipline has traditionally been centered on the human translator and her individual competence, the notion of "translation quality", in translation studies, has in the last decades assumed a multi-faceted shape, embracing aspects that go beyond an individual's competence of optimising the relation between linguistic naturalness and semantic fidelity or her ability to use rule sets specific to a given language pair.

This paper presents a study on human and automatic evaluations of translations in the French-German translation learner corpus KOPTE (Wurm, 2013). The aim of the paper is to shed light on the differences between MT evaluation scores and approaches to translation evaluation rooted translation studies. We illustrate the factors contributing to the human evaluation of translations, opposing these factors to the results of automatic evaluation metrics, by applying two of the most popular automatic evaluation metrics, namely BLEU (Papineni et al., 2002) and Meteor (Denkowski and Lavie, 2011), to a sample of human translations available from KOPTE. The goal of these experiments is threefold. Firstly, we want to study whether the automatic scores can mimic the fine-grained distinctions of the human translations expert who evaluated the translations available from KOPTE or, at least, make meaningful distinctions when applied to human translations. Secondly, we are interested in investigating how automatic evaluation scores evolve if the number of chosen references is increased. Finally, we are also interested in examining whether a higher number of references influence the correlation of the automatic scores with the human expert grades for the same translation. Our experiments suggest that both BLEU and Meteor systematically underestimate the quality of the translations tested.

By means of a qualitative analysis of human translations we then highlight the concept of legitimate variation and attempt to reveal weaknesses of automatic evaluation metrics. More specifically, our qualitative analysis suggests that lexical similarity scores are neither to cope satisfactorily with standard lexical variation (paraphrases, synonymy) nor with dissimilarities that can be traced back to the source text or the nature of the translation process itself.

References

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*.

Kishore Papineni, Salim Roukos, ToddWard, andWei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pages 311–318.

Andrea Wurm. 2013. Eigennamen und Realia in einem Korpus studentischer Übersetzungen (KOPTE). *transkom*, 6:381–419, 2

Assessing Inter-Annotator Agreement for Translation Error Annotation

Arle Lommel, Maja Popović, Aljoscha Burchardt
E-mail: arle.lommel@dfki.de, maja.popovic@dfki.de, aljoscha.burchardt@dfki.de

One of the key requirements for demonstrating the validity and reliability of an assessment method is that annotators be able to apply it consistently. Automatic measures such as BLEU traditionally used to assess the quality of machine translation gain reliability by using human-generated reference translations under the assumption that mechanical similar to references is a valid measure of translation quality. Our experience with using detailed, in-line human-generated quality annotations as part of the QTLaunchPad project, however, shows that inter-annotator agreement (IAA) is relatively low, in part because humans differ in their understanding of quality problems, their causes, and the ways to fix them. This paper explores some of the facts that contribute to low IAA and suggests that these problems, rather than being a product of the specific annotation task, are likely to be endemic (although covert) in quality evaluation for both machine and human translation. Thus disagreement between annotators can help provide insight into how quality is understood.

Our examination found a number of factors that impact human identification and classification of errors. Particularly salient among these issues were: (1) disagreement as to the precise spans that contain an error; (2) errors whose categorization is unclear or ambiguous (i.e., ones where more than one issue type may apply), including those that can be described at different levels in the taxonomy of error classes used; (3) differences of opinion about whether something is or is not an error or how severe it is. These problems have helped us gain insight into how humans approach the error annotation process and have now resulted in changes to the instructions for annotators and the inclusion of improved decision-making tools with those instructions. Despite these improvements, however, we anticipate that issues resulting in disagreement between annotators will remain and are inherent in the quality assessment task.

Talks Session 2

11:00 – 13:00
Chairpeople: KeithMiller and Stacey Bailey

TURKOISE: A Mechanical Turk-based Tailor-made Metric for Spoken Language Translation Systems in the Medical Domain

Marianne Starlander
E-mail: Marianne.Starlander@unige.ch

In this paper, we will focus on the evaluation of MedSLT, a medium-vocabulary hybrid speech translation system intended to support medical diagnosis dialogues between a physician and a patient who do not share a common language. How can the developers ensure a good quality to their users, in a domain where reliability is of the highest importance?

MedSLT was designed with a strong focus on reliability in the correct transmission of the message. One of the characteristics of MedSLT is its rule-based architecture that uses an interlingua approach to produce highly reliable output. This approach avoids surface divergences in order to keep only the meaning of the sentences. Consequently, sentences are translated more freely and as a consequence of our speech input, the sentences are particularly short. These characteristics entail quite low BLEU scores as well as little correlation with human judgment. Besides these automatic metrics, we also completed several human evaluations; using different scales (including fluency and adequacy as well as ranking). None of our experimented metrics gave us satisfactory results in the search of an operational metric for speech translation systems in a safety-critical domain such as the medical diagnosis domain. We have thus decided to experiment with manual metrics in order to find

an evaluation that could be implemented without producing human references and at reasonable cost, within a minimum time span.

In the following paper we will describe the path that led us to using Amazon Mechanical Turk[1] (AMT) as an alternative to more classical automatic or human evaluation. We started using adequacy and fluency metrics but soon decided to experiment with a tailor-made and task-specific human metric, adapted to our domain but that could be used by a wider group of evaluators thanks to the AMT while guaranteeing certain coherence between the evaluators. The proposed metric is called **TURK**OISE, designed to be used by unskilled AMT evaluators while guaranteeing reasonable level of coherence between evaluators.

Our study focuses on inter-rater agreement comparing this aspect for our in-house small group of translator-evaluators compared to a wider group of AMT workers. We would also like to quantify the effort in running the AMT evaluation in order to compare the resources needed. Developers and researchers tend to minimize the effort related with the creation of reference translations in order to use BLEU or other reference-based metrics. Hence, we assume that if AMT workers are found to be reliable, this type of evaluation would be, at least, as cost and time effective as the classical automatic metrics but providing the advantage of reflecting the end-user's quality level request.

Our main results of this experiment are that AMT workers are found to be reaching comparable levels of inter-rater agreement when using the classic fluency and adequacy metrics, but also TURKoise, being our tailor-made evaluation scale.

MADCAT Evaluation: Operational Accuracy of MT Applied to OCR

Caitlin Christianson, Bonnie Dorr, Joseph Olive
E-mail: {caitlin.christianson.ctr,joseph.olive.ctr}@darpa.mil, bdorr@ihmc.us

Evaluating progress is an important aspect of NLP research that can help identify the most effective techniques and systems. Prior evaluation techniques have been valuable in monitoring progress and comparing different MT systems, but they have failed to answer an important question relevant to research sponsors with an interest in operational MT use, namely what accuracy is necessary for any given application. To answer this question, we devised an experiment to solicit input from experienced users of translated material by providing them documents with varying levels of translation accuracy and asking them which of these documents would be useful for a given task. We were interested mainly in three tasks: editing, gisting, and triage. Documents deemed editable would be publishable with human editing, documents deemed gistable would be suitable for human readers to determine the basic meaning, and documents deemed triageable would be suitable for determining mission relevance.

Current MT algorithms were used to translate Arabic and Spanish documents, and accurate human translations were obtained. The MT outputs were then edited to reflect the meaning of the human translated documents. Errors were counted (insertions, deletions, substitutions and moves of any number of adjacent words) and divided by the number of words in the source document. Both machine-translated documents had, on average, 45% errors. The MT output was then corrected by randomly choosing the edited corrections in steps of 5% to generate 10 documents. The original MT output and human translations were added to these for a total of 12 documents. The results showed that triageable, gistable and editable documents required accuracy of 55%, 70%, and 85%, respectively. This work led to a new evaluation paradigm, Human-mediated Translation Error Rate (HTER; Olive and Christianson, 2011; Dorr et al., 2011). This meaning-based metric compares machine-translated text to a "gold standard" translation of the same text created by a team of human

[1] www.mturk.com

translators. The MT output is edited to obtain text that conveys the same meaning as the "gold standard" text; the number of edits are counted and divided by the number of words.

Our research in applying HTER has focused on the use of various evaluation methods to determine MT accuracy in relation to technology applications. An OCR-based example of this relation was investigated for the Multilingual Automatic Document Classification, Analysis, and Translation (MADCAT) program. An Arabic data set was generated to determine for measuring translation accuracy. However, to ascertain the applicability of MADCAT systems on operational data, the program acquired some hand-written documents collected in the field in Iraq. Consistent improvements in the ability of MADCAT to translate program-generated documents were obtained during the first 4 years of the project. However, for the field-collected documents, the starting point was much lower – not even triageable. By year 5, improvements were still significant – above gistable.

We have made the case for evaluation in context, using HTER on the final MT output, rather than standard transcription error rates used in OCR. We have also argued for determining performance levels on data with operational characteristics and relating accuracy judgments to utility levels of relevance to the end user.

References

Dorr, B.; Olive, J; McCary, J.; Christianson, C. (2011) "Chapter 5: Machine Translation Evaluation and Optimization," in Olive, J; Christianson, C; McCary, J. (Eds.), *Handbook of NLP and MT: DARPA Global Autonomous Language Exploitation*, pp. 745—843.

Olive, J; Christianson, C. (2011) "The GALE Program," in Olive, J; Christianson, C; McCary, J. (Eds.), *Handbook of NLP and MT: DARPA Global Autonomous Language Exploitation*, pp. vii—xiv.

Continuous Operational Evaluation of Evolving Proprietary MT Solution's Translation Adequacy

Ekaterina Stambolieva
E-mail: Ekaterina.stambolieva@euroscript.lu

Little attention is given to the focus on continuous diagnostic monitoring of the adequacy of translations (Koehn, 2010) dependent on specific business scenarios. Numerous organizations, including ours, post-edit Machine Translation (MT) to accelerate translation time-to-delivery and reduce translation costs. Unfortunately, in many cases, MT quality is not good enough for the task of post-editing and MT systems struggle to deliver native-fluency translations (Allen, 2003). Many researchers (Krings 2001, He et al. 2010, Denkowski and Lavie 2012, Moorkens and O'Brien 2013) agree that human end-user (translators, project coordinators with solid linguistic and translation knowledge, among othres) evaluation input contributes to MT quality improvement. Armed with translators' feedback, benchmarks and metrics such as QTLaunchPad[2]'s MQM (Doherty et al., 2013) along with taraXÜ2[3]'s confidence score (Avramidis et al., 2011) tackle the MT quality problem in search of effective quality evaluation. Nevertheless, all of these do not solve the impending industry problem – evaluating and comparing over-time MT solution modifications. This paper contributes to the development of a Continuous Operational MT Evaluation (COMTE) approach, which concentrates on repeated evaluation of MT system improvements based on human end-user feedback.

It is crucial to secure high MT adequacy on each stage of the MT system modifications that reflect the human translators' assessment and expectation of the output. COMTE contributes to quality improvement in modified MT solutions based on end-users' feedback, and helps to increase post-

[2] http://www.qt21.eu/launchpad/
[3] http://taraxu.dfki.de/

editing task suitability. COMTE does not directly evaluate translation quality, like scores such as BLEU (Papineni et al., 2002) and METEOR (Banerjee et al., 2005), or metrics such as MQM. Instead COMTE assesses the over-time MT system improvement. It focuses on measuring translation adequacy and fluency based on developments, which solve MT solution issues and are suggested by the system's end-users. We propose to measure continuously translation adequacy in the task of post-editing by employing two well-known evaluation metrics. We explore the correlation between the metrics and the scheduled human-evaluation-driven MT system modifications.

The two founding scores of the approach are: Edit Distance (ED) (Przybocki et al., 2006) and Fuzzy Match (FM) (Bowker, 2002). ED measures in how many edits machine translation output transforms into a human translated segment. ED is employed as a simple metric that measures post-editing effort. On the other hand, FM is a metric inherited by computer-assisted translation (CAT) tools. It shows what percentage of the current text for translation can be fluently translated by selecting an existing translation from business-dependent Translation Memories (TM). In the language business, a FM threshold is set, on which many pricing strategies depend. All text that has a FM lower than the fixed threshold is machine translated. A TM match is retrieved for the rest. Importantly, this approach requires zero additional annotation effort – all the information is derived from collected structured translators' feedback of the MT output. We also show how ED and FM correlate depending on the business scenario and MT quality improvements.

Our approach suggests a reliable strategy for performing operational translation evaluation of evolving in-house MT systems with wide applicability in the language industry. The evolution of the systems is based on human end-user feedback collected in a systematic way, following a 4-category error typology. We present empirical evidence that ED and FM correlate with successful system improvements, and conclude that they can thus be used to automatically assess system development.

References

Jeffrey Allen. 2003. Post-editing. In Harold Somers, editor, *Computers and Translation: A Translator's Guide*, pages 297-317, John Benjamins B.V.

Eleftherios Avramidis, Maja Popovic, David Vilar Torres and Aljoscha Burchardt. 2011. Evaluate with Confidence Estimation: Machine ranking of translation outputs using grammatical features. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT-11)*, Edinburgh, United Kingdom.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Matric for MT Evaluation with Improved Correlation with Human Judgements. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguists (ACL-2005)*, Ann Arbor, Michigan.

Lynne Bowker. 2002. *Computer-Aided Translation Technology: A practical Introduction*, pages 98-101, University of Ottawa Press.

Michael Denkowski and Alon Lavie. 2012. Challenges in Predicting Machine Translation Utility for Human Post-Editors. In *Proceedings of ATMA 2012*, San Diego.

Stephen Doherty, Federico Gaspari, Declan Groves, Josef van Genabith, Lucia Specia, Aljoscha Burchardt, Arle Lommel and Hans Uszkoreit. 2013. *Mapping the Industry I: Findings on Translation Technogogies and Quality Assessment*. European Commission Report.

Yifan He, Yanjun Ma, Johann Roturier, Andy Way and Josef van Genabith 2010. Improving the post-editing experience using translation recommendation: A user study. In *Proceedings of the 9th Annual AMTA Conference*, pages 247-256, Denver, CO.

Phillip Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, pages 217-218.

Chunk Accuracy: A Simple, Flexible Metric for Translation Quality

Lars Ahrenberg
E-mail: lars.ahrenberg@liu.se

Many approaches to assessment of translations are based on error counts. These are usually supported by detailed taxonomies that highlight different quality aspects. Even if provided with guidelines the categorization and localization of errors can be quite difficult and time-consuming for a human annotator. Efforts may be wasted on details that are not really relevant for the purpose at hand. For example, a post-editor may be more helped by getting information on the locations of errors than a detailed classification of them.

A framework such as MQM: Multidimensional Quality Metrics (Uszkoreit&Lommel, 2013) is very helpful as a guide to what may be relevant for a given evaluation purpose. There is still a problem of applying criteria, however, once you have a taxonomy. Even if your selection is small, it is still often multi-dimensional, and ambiguities are likely to arise. For example, the distinction between error categories such as Wrong Translation and Missing Word may be clear in principle, but can be hard to make in a concrete case. Also, the question remains how a multi-dimensional selection is used to compare systems. As Williams (2001: 329) puts it: "The problem is this: assuming you can make a fair assessment of each parameter, how do you then generate an overall quality rating for the translation?"

I suggest that these two problems can be at least partly remedied by the following measures: (1) use the simplest possible taxonomies and give priority to counts before types; (2) use chunks as the loci of problems; a chunk can be read as a single unit by a human and eases the task of assigning a problem to a particular word, as for instance in the case of agreement errors. Still, it is more informative than counting errors for the whole text or complete sentences. For example, a post-editor may be shown not just that there are errors in a sentence, but in which part of the sentence the errors are located.

In the simplest case chunks need only be categorized into problematic (P) and correct (C). The metric then becomes C/(C+P) (or a percentage). To increase granularity, we can use a n-ary scale (for example good, bad, and ugly as is currently popular) and report a distribution over these categories. To get more informative we can categorize problems as those pertaining to adequacy (relation to corresponding source chunks), fluency (target language problems) and others. And then climb further down a taxonomy such as the MQM as motivated by the evaluation purpose.

Chunk accuracy can be applicable whenever a micro-level analysis is called for, e.g., in assessment of student translations, in post-editing settings, or even for MT development. It can be automated to a some extent, thus reducing the human effort. While aggregating observations at the micro-level, and reporting quality characteristics, it is not a final assessment, however. It reports values, not thresholds.

In my presentation, I will further develop my arguments and make comparisons of chunk accuracy to other known frameworks for error analysis.

References

H. Uszkoreit and A. Lommel (2013). Multidimensional Quality Metrics: A New Unified Paradigm for Human and Machine Translation Quality Assessment.(http://www.qt21.eu/launchpad/sites/default/files/MQM.pdf)

M. Williams (2001). The Application of Argumentation Theory to Translation Quality Assessment. Meta 46(2): 326-344.

Word Transition Entropy as an Indicator for Expected Machine Translation Quality

Michael Carl and Moritz Schaeffer
Email: mc.ibc@cbs.dk, moritzschaeffer@gmail.com

While most machine translation evaluation techniques (BLEU, NIST, TER, METEOR) assess translation quality based on a single (or a set of) reference translations, we suggest to evaluate the literality of a set of (human or machine generated) translations to infer their potential quality. We provide evidence which suggests that more literal translations are produced more easily, by humans and machine, and are also less error prone. Literal translations may not be appropriate or even possible for all languages, types of texts, and translation purposes. However, in this paper we show that an assessment of the literality of translations allows us to (1) evaluate human and machine translations in a similar fashion and (2) may be instrumental to predict machine translation quality scores.

While "translators tend to proceed from more literal versions to less literal ones" (Chesterman, 2011) it is controversial what it actually means for a translation to be literal. In this paper, we follow a strict definition which defines literal translations to "consist of the same number of lexical words, representing equivalent grammatical categories, arranged in the same literal order and underlying semantically equivalent sentences" (Krzeszowski, 1990). This definition is operationalized by the following criteria:

1. Word order is identical in the source and target languages
2. Source and target text items correspond one-to-one
3. Each source word has only one possible translated form in a given context.

In this talk we focus on point 3: Nine English source texts were machine-translated into Spanish and post-edited by nine different post-editors. The post-edited texts were subsequently corrected by independent reviewers. The translations were semi-automatically aligned on a sentence and a word level. Keystroke and gaze data was collected during a part of the post-editing sessions. See Carl et al, (2014) for a more detailed description of the experimental data.

We computed the edit distance between the MT output and the post-edited text (*MT-PE*) and the edit distance between the post-edited translation and its reviewed version (*PE-RE*). We take the *MT-PE* distance as a quality indicator for the MT output: the more a post-editor modifies the MT output the worse can be expected the MT quality to be and the bigger will be the *MT-PE* distance.

We computed the word translation entropy of the human post-edited texts *HH(e):* the word translation probabilities $p(e \rightarrow si)$ of an English word e into the Spanish word *si* were computed as the ratio of the number of alignments e-si in the post-edited texts. Subsequently, the entropy of an English source word *e* was computed as:

$$(1) \qquad HH(e) = -1 * \Sigma i\ p(e \rightarrow si) * log2(p(e \rightarrow si)).$$

We also computed the word transition entropy in the machine translation search graph MH(e) based on the transition probabilities $p(e \rightarrow si)$ from *s-1* to *si* as provided in the Moses search graphs. Up to 10% of the most unlikely transitions were discarded, and transition weights of the remaining translation options were mapped into probabilities.

Given these metrics, we can make a number of assumptions: If the MT output of a source word e was not modified by any post-editor, then $HH(e)=0$. Conversely, *HH(e)* would reach its maximum value if the MT output for e was modified by every post-editor in a different way. If a segment was not modified at all, *MT-PE* would be 0 and hence we expect a positive correlation between *HH(e)* and *MT-PE*.

Further, we expect that translations become worse as the entropy *MH(e)* increases, since it might be more difficult for an MT system to decide which translation to choose if several word transition probabilities are similarly likely, and thus the likelihood may increase for a sub-optimal translation choice. Subsequently, we expect to see more post-editing activities on translations with higher *MH(e)* values, and thus a positive correlation between *HH(e)* and *MH(e)*. Finally, we expect that textual changes of the MT output require more gaze and translation time, so that we also expect a positive correlation between post-editing activities and *MH(e).* In our analysis we show that:

1. *HH(e)* correlates positively with the edit distance *MT-PE*. That is, the edit distance increases if different post-editors translate a source word *e* in different ways.
2. There is a negative correlation between *MT-PE* and *PE-RE*: the more a text was edited in the post-edited phase, the less it was modified during revision, and vice versa.
3. *HH(e)* correlates with the gaze duration on (and translation production time of) *e*. That is, it is more time consuming for a translator to translate a source language word which can be translated in many different ways, than a source word which translates only into few target words, with high probability.
4. *HH(e)* correlates with *MH(e).* That is, if human translators translate a source word in many different ways, also the SMT system has many translation options for that word.

In this paper we pinpoint a correlation between the entropy of human translation realizations and the entropy of machine translation representations. As such, this is not surprising, since statistical machine translation systems are trained on, and thus imitate, the variety of human produced translations. Entropy is tightly linked to translation literality, and as translations become less literal (be it for structural reasons or for translator's choices) state-of-the-art statistical machine translation systems fail, while human translators seem to deploy as of now non-formalized translation strategies, to select amongst the many possible the/a good translation. This turning point may serve as an indicator for translation confidence, beyond which the quality of MT output becomes less reliable, and thus MT post-editing may become less effective.

References

Michael Carl, Mercedes Martínez García, Bartolomé Mesa-Lao, Nancy Underwood (2014) "CFT13: A new resource for research into the post-editing process". In *Proceedings of LREC*.

Chestermann, Andrew. (2011) Reflections on the literal translation hypothesis. In *Methods and Strategies of Process Research*, edited by Cecilia Alvstan, Adelina Held and Elisabeth Tisselius, Benjamins Translation Library, pp. 13-23, 2011.

Krzeszowski, Tomasz P. (1990) Contrasting Languages: The Scope of Contrastive Linguistics. Trends in Linguistics, Studies and Monographs, Mouton de Gruyter.

A New Multiple Choice Comprehension Test for MT

Douglas Jones, Paul Gatewood, Martha Herzog, Tamas Marius
E-mail: daj@ll.mit.edu, paul.gatewood@ll.mit.edu, MHerzog2005@comcast.net, tamas.marius@dliflc.edu

We present results from a new machine translation comprehension test, similar to those developed in previous work (Jones et al., 2007). This test has documents in four conditions: (1) original English documents; (2) human translations of the documents into Arabic; conditions (3) and (4) are machine translations of the Arabic documents into English from two different MT systems. We created two forms of the test: Form A has the original English documents and output from the two Arabic-to-English MT systems. Form B has English, Arabic, and one of the MT system outputs. We administered the comprehension test to three subject types recruited in the greater Boston area: (1) native English speakers with no Arabic skills, (2) Arabic language learners, and (3) Native Arabic speakers who also have English language skills. There were 36 native English speakers, 13 Arabic learners, and 11 native Arabic speakers with English skills. Subjects needed an average of 3.8 hours to complete the test, which had 191 questions and 59 documents. Native English speakers with no Arabic skills saw Form A. Arabic learners and native Arabic speakers saw form B.

The overall comprehension results for English natives reading Form A were 88% for the original English texts, 53% for MT1, and 76% for MT2. System level BLEU scores were 0.0624 and 0.2158 respectively. Comprehension scores were not strongly correlated with BLEU scores. For the Arabic language learners who saw Form B, the comprehension results were 91% for English, 46% for Arabic, and 76% for MT2. For the Arabic native speakers who saw Form B, comprehension results were 82% for English, 80% for Arabic, and 72% for MT2. The Arabic learners, who had an average of 2.5 semesters of Arabic instruction at the college level, demonstrated comprehension at a level between that of MT1 and MT2 as read by native English speakers. No MT results were as good as native speakers reading their native language.

We used the standardized language skill descriptions defined by the Interagency Language Roundtable (ILR); see (ILR, 2014). To measure machine translation capabilities, as opposed to human foreign language capabilities, we constructed a variant of the Defense Language Proficiency Test, following the general DLPT design principles, but modified to measure the quality of machine translation. This test is a multiple-choice format, ILR-based machine translation test format described in the paper "ILR-Based MT Comprehension Test with Multi-Level Questions" by Jones et al. in the proceedings of HLT 2007. Test documents were rated for ILR reading skills and were split between Levels 2, 2+ and 3. Questions were also rated for ILR level: Level 1, 2, 2+, and 3; comprehension results generally reflected the difficulty levels.

References

ILR. (2014). Interagency Language Roundtable (website). http://www.govtilr.org/.

Jones, D., et al (2007). ILR-Based MT Comprehension Test with Multi-Level Questions. In Human Language Technologies 2007: NAACL, Short Papers, pages 77–80.

This work is sponsored by the Defense Language Institute under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Standardized ILR-Based and Task-Based Speech-to-Speech MT Evaluation

Douglas Jones, Paul Gatewood, Martha Herzog, Tamas Marius
E-mail: daj@ll.mit.edu, paul.gatewood@ll.mit.edu, MHerzog2005@comcast.net, tamas.marius@dliflc.edu

This paper describes a new method for task-based speech-to-speech machine translation evaluation, in which tasks are defined and assessed according to independent published standards, both for the military tasks performed and for the foreign language skill levels used. We analyze task success rates and automatic MT evaluation scores for 220 role-play dialogs. Each role-play team consisted of one native English-speaking soldier role player, one native Pashto-speaking local national role player, and one Pashto/English interpreter. The overall PASS score, averaged over all of the MT dialogs, was 44%. The average PASS rate for HT was 95%.

Scenarios were of two general types: a basic definition without any complications, and a contrasting definition with some type of obstacle, perhaps minor, that needed to be overcome in the communication. For example, in a basic Base Security scenario, a Local National may seek permission to pass a checkpoint with valid identification. In a contrast scenario, he may lack the identification, but seek an alternative goal that does not involve passing the checkpoint. Overall PASS/FAIL results for the HT condition were 95% for basic scenarios and 94% for contrasting scenarios with obstacles. For MT we observed 67% PASS for basic and 35% for contrast scenarios. The performance gap between HT at 94~95% and MT with basic scenarios at 67% was 27% on average, whereas the difference between MT in basic scenarios and MT in contrasting scenarios was 32%.

The dialogs were also assessed for language complexity. Scenarios with language complexity at the ILR Levels 1, 1+ and 2 had PASS scores of 94%, 100% and 92% respectively in the HT condition. For MT the overall results were 47%, 48% and 31%. In other words, MT does not work as well when the language is fundamentally more complex. The average BLEU score for English-to-Pashto MT was 0.1011; for Pashto-to-English it was 0.1505. BLEU scores varied widely across the dialogs. Scenario PASS/FAIL performance was also not uniform within each domain. Base Security scenarios did perform relatively well overall. Some of the scenarios in other domains were performed well with MT but performance was uneven.

Role players performed 20 tasks in 4 domains. The domain-level PASS scores ranged from 89% to 100% in the HT condition. For MT we observed 83% PASS rate in one domain, Base Security, with the remaining three domains ranging from 26% to 50%. The dialogs were human-scored in two main ways: (a) aggregate PASS/FAIL outcomes, and (b) a secondary assessment of specific communication initiatives. Inter-coder agreement for task PASS/FAIL scoring, which required an assessment of several performance measures per task, averaged 83%. Agreement for the specific communication initiatives was 98%.

We learned that success rates depended as much on task simplicity as it did upon the translation condition: 67% of simple, base-case scenarios were successfully completed using MT, whereas only 35% of contrasting scenarios with even minor obstacles received passing scores. We observed that MT had the greatest chance of success when the task was simple and the language complexity needs were low.

Rating Evaluation Methods through Correlation

Lena Marg
E-mail: lena.marg@welocalize.com

While less attention may have been dedicated to operational or task-based evaluation metrics by the MT research community, in our case (i.e. Language Service Provider), every evaluation is by definition task-based as it is carried out at the request of or tailored to a specific end-client and therefore with a defined purpose, scope and end-user in mind. This being the case, there still seems to be a need for more appropriate and easy-to-use metrics, both for evaluating the quality of raw and post-edited MT versus human translation.

In 2013, we put together a database of all evaluations (automatic scorings, human evaluations including error categorization and productivity tests including final quality assessments) carried out that year in Welocalize, in order to establish correlations between the various evaluation approaches, draw conclusions on predicting productivity gains and also to identify shortcomings in evaluation approaches. The database was limited to evaluations of that year for consistency in approach with regard to human evaluations and productivity tests compared to previous years.

Among the findings we observed were that the Human Evaluations of raw MT (especially the "Accuracy" score) seemed to be stronger predictors for potential productivity gains than automatic scores; Human Evaluation error categorizations provided initial glimpses of (cognitive effort) trends, but the markings seemed to be unreliable to some extent; further analysis, adjustment and fine-tuning of the (final) QA process are needed.

As part of the workshop, I would like to share findings from our data correlation analysis, which metrics turned out to be most valid and where we identified shortcomings. I will also be able to share first steps taken to improve our evaluation protocols in ongoing tests.

Description of metrics used in correlation database

The automatic score used for the data correlation is BLEU. When produced by an MT system, it would be based on MT versus human reference from a TM. In the case of productivity tests they can also be generated from MT versus post-edited version of the given content.

Human Evaluations of raw MT output are scored on a scale from 1-5, with 5 indicating "very good" quality and 1 indicating "very low" quality. They are divided into three parts: Accuracy score, Fluency score and Error Categorization. Human Evaluations are typically carried out on a set of manually selected strings representative of the content to be evaluated (i.e.: string length; typical "pitfalls" such as handling of software options, measurements and conversions, "To"-structures, gerunds, marketing speak, enumerations, elliptical structures etc.).

Productivity Tests are carried out in iOmegaT, an instrumented version of the open-source CAT tool co-developed by John Moran and Welocalize, which captures the translation time and number of edits for each segment. Test kits contain a mix of segments to translate from scratch and segments to post-edit, and linguists are usually asked to carry out 8 hours of translation/post-editing work.

Similar to the LISA QA Model and SAE J2450, the current QA metrics are a quantitative-based method of translation quality assessment which measures the number, severity and type of errors found in a text and calculates a score, which is indicative of the quality of a given translation.

Relating Translation Quality Barriers to Source-Text Properties

Federico Gaspari, Antonio Toral, Arle Lommel,
Stephen Doherty, Josef van Genabith, Andy Way
E-mail: {fgaspari, atoral, away}@computing.dcu.ie, arle.lommel@dfki.de,
s.doherty@unsw.edu.au, josef.van_genabith@dfki.de

This study presents work on the identification of translation quality barriers. Given the widely perceived need to enhance MT quality and the reliability of MT evaluation for real-life applications, this study is of potential interest to a variety of MT users and developers. Our study focuses on identifying the source-side linguistic properties that pose MT quality barriers for specific types of MT systems (statistical, rule-based and hybrid) and for output representative of different quality levels (poor-, medium- and high-quality) in four translation combinations, considering English to and from Spanish and German. Using the diagnostic MT evaluation toolkit DELiC4MT and a set of human reference translations, we relate translation quality barriers to a selection of 9 source-side PoS-based linguistic checkpoints (adjectives, adverbs, determiners, common nouns, nouns, proper nouns, particles, pronouns and verbs).

DELiC4MT is an open-source toolkit for diagnostic MT evaluation. Its diagnostic dimension derives from its ability to focus on user-defined linguistic checkpoints, i.e. phenomena of the source language that the user decides to analyse when evaluating the quality of MT output. Linguistic checkpoints can correspond to interesting or difficult lexical items and/or grammatical constructions for which a specific translation quality assessment is required. They can be defined at any level of granularity desired by the user, considering lexical, morphological, syntactic and/or semantic information.

DELiC4MT has so far been used to evaluate the overall quality of MT systems with respect to their performance on user-defined source-side linguistic phenomena. The novelty of this work lies in the application of this toolkit to the investigation of translation quality barriers. These are investigated according to two main variables. Firstly, we consider different MT system types: this variable enables us to compare the performance of statistical, rule-based and hybrid MT software on a selection of source-language linguistic checkpoints. Secondly, we look at human quality rankings of the MT output: this variable concerns the quality band assigned by human evaluators to the output of each MT system, whereby each sentence was rated as either good (rank 1), near-miss (rank 2) or poor (rank 3). We are thus able to evaluate the performance of the MT systems on each checkpoint separately for those sentences that fall into each of these rating bands.

We show that the combination of manual quality ranking and automatic diagnostic evaluation on a set of PoS-based linguistic checkpoints is able to identify the specific quality barriers of different MT system types across the four translation directions under consideration. On the basis of this evaluation, we have analysed the correlation between the scores obtained for each of these source-side linguistic phenomena and the human quality ratings, thus assessing the extent to which these phenomena can be used to predict human quality evaluation. Considering all the MT system types evaluated together, it turns out that the best predictors are verbs (r=0.795), proper nouns (r=0.658) and pronouns (r=0.604), while the worst one is by far adverbs (r=0.02).

Keywords: MT quality barriers, diagnostic evaluation, statistical/rule-based/hybrid MT, linguistic features

Hands-on Exercises

14:00 – 18:00
Chairperson: Kim Harris

Participants will gain hands-on experience with many of the evaluation methodologies presented during the morning sessions. Discussion and comparison of methodologies will follow.

Language Resources and Technologies for Processing and Linking Historical Documents and Archives- Deploying Linked Open Data in Cultural Heritage – LRT4HDA

26 May 2014

ABSTRACTS

Editors:

Kristín Bjarnadóttir, Mathew Driscoll, Steven Krauwer, Stelios Piperidis, Cristina Vertan, Martin Wynne

Workshop Programme

09:00 - 09:30 – Opening and introduction by Workshop Chairs*]*

09:30- 10:30 – Invited Talk
Eiríkur Rögnvaldsson, *Old languages, new technologies: The case of Icelandic*

10:30 – 11:00 Coffee break

11:00 – 11:30 – Session Language Resources
Patrick Schone, *A Personal Name Treebank and Name Parser to Support Searching and Matching of People Names in Historical and Multilingual Contexts*

11:30 – 12:00 – Session Language Resources
Elaine Uí Dhonnchadha, Kevin Scannell, Ruairí Ó hUiginn, Eilís Ní Mhearraí, Máire Nic Mhaoláin, Brian Ó Raghallaigh, Gregory Toner, Séamus Mac Mathúna, Déirdre D'Auria, Eithne Ní Ghallchobhair and Niall O'Leary, *Corpas na Gaeilge (1882-1926): Integrating Historical and Modern Irish Text*

12:00 – 12:30 – Session Language Resources
Ásta Svavarsdóttir, Sigrún Helgadóttir and Guðrún Kvaran, *Language resources for early Modern Icelandic*

12:30 – 12:50 – Session Language Resources
Dominique Ritze, Caecilia Zirn, Colin Greenstreet, Kai Eckert and Simone Paolo Ponzetto, *Named Entities in Court: The MarineLives Corpus*

12:50 – 14:00 Lunch break

14:00 – 14:20 – Session Language Resources
Stephen Tyndall, Building *Less Fragmentary Cuneiform Corpora: Challenges and Steps Toward Solutions*

14:20 – 14:40 – Session Language Resources
Thorhallur Eythorsson, Bjarki Karlsson and Sigríður Sæunn Sigurðardóttir, *Greinir skáldskapar*: *A diachronic corpus of Icelandic poetic texts*

14:40 – 15:00 – Session Historical Newspaper Archives
Oliver Pfefferkorn and Peter Fankhauser, *On the Role of Historical Newspapers in Disseminating Foreign Words in German*

15:00 – 15:20 – Session Historical Newspaper Archives
Örn Hrafnkelsson and Jökull Sævarsson, *Digital libraries of historical Icelandic newspapers, periodicals, magazines and old printed books*

15:20 – 15:50 – Session Historical Newspaper Archives
Susanne Haaf and Matthias Schulz, *Historical Newspapers & Journals for the DTA*

16:00 – 16:30 Coffee break

16:30 – 17:00 – Session Tools for analysis of historical documents
Jón Daðason, Kristín Bjarnadóttir and Kristján Rúnarsson, *The Journal Fjölnir for Everyone: The Post-Processing of Historical OCR Texts*

17:00 – 17:20 – Session Tools for analysis of historical documents
Ludger Zeevaert, *IceTagging the "Golden Codex". Using language tools developed for Modern Icelandic on a corpus of Old Norse manuscripts*

17:20 – 17:40 – Session Tools for analysis of historical documents
Federico Boschetti, Andrea Cimino, Felice Dell'Orletta, Gianluca Lebani, Lucia Passaro, Paolo Picchi, Giulia Venturi, Simonetta Montemagni and Alessandro Lenci, *Computational Analysis of Historical Documents: An Application to Italian War Bulletins in World War I and II*

17:40 – 18:00 – Session Tools for analysis of historical documents
Cristina Vertan, Walther v. Hahn, *Discovering and Explaining Knowledge in Multilingual Historical Documents*

18:00 – 18:30 Discussions and Closing

Workshop Organizers

Kristín Bjarnadóttir	The Arni Magnusson Institute for Icelandic Studies, Iceland Arnamagnean Commission, Copenhagen,
Mathew Driscoll	Denmark
Steven Krauwer	CLARIN ERIC, Netherlands
Stelios Piperidis	ILSP, Athens, Greece
Cristina Vertan	University of Hamburg
Martin Wynne	University of Oxford, UK

Workshop Programme Committee

Lars Borin	University of Gothenburg, Sweden
Rafael Carrasco	University of Alicante, Spain
Paul Doorenbosch	National Library of the Netherlands, Netherlands
Thorhallur Eythorsson	University of Iceland, Iceland
Alexander Geyken	BBAW, Germany
Günther Görz	University Erlangen, Germany
Walther v. Hahn	University of Hamburg, Germany
Erhard Hinrichs	University of Tuebingen, Germany
Guillaume Jacquet	JRC, Italy
Marc Kupietz	IDS, Germany
Éric Laporte	Université Paris-Est Marne-la-Vallée, France
Piroska Lendvai	Hungarian Academy of Sciences, Hungary
Thierry Paquet	LITIS, France
Gábor Prószéky	MorphoLogic, Hungary
Bente Maegaard	University of Copenhagen, Denmark
Christian Emil Ore	University of Oslo, Norway
Eiríkur Rögnvaldsson	University of Iceland, Iceland
Petya Osenova	IICT, Bulgarian Academy of Sciences, Bulgaria
Manfred Thaller	Cologne University, Germany
Tamás Váradi	Hungarian Academy of Sciences, Hungary
Matthew Whelpton	University of Iceland, Iceland
Kalliopi Zervanou	University of Tilburg, the Netherlands

Foreword

Recently, the collaboration between the NLP community and the specialists in various areas of the Humanities has become more efficient and fruitful due to the common aim of exploring and preserving cultural heritage data. It is worth mentioning the efforts made during the digitisation campaigns in the last years and within a series of initiatives in the Digital Humanities, especially in making Old Manuscripts available through Digital Libraries.

Having in mind the number of contemporary languages and their historical variants, it is practically impossible to develop brand new language resources and tools for processing older texts. Therefore, the real challenge is to adapt existing language resources and tools, as well as to provide (where necessary) training material in the form of corpora or lexicons for a certain period of time in history.

Another issue regarding historical documents is their usage after they are stored in digital libraries. Historical documents are not only browsed but together with adequate tools they may serve as basis for re-interpretation of historical facts, discovery of new connections, causal relations between events etc. In order to be able to make such analysis, historical documents should be linked among themselves, on the one hand, and with modern knowledge bases, on the other. Activities in the area of Linked Open Data (LOD) play a major role in this respect.

A particular type of historical documents are the newspaper collections and archives. Newspapers reflect what is going on in society, and constitute a rich data collection for many types of humanities research, ranging from history, political and social sciences to linguistics, both synchronic and diachronic, and both national and cross-national. They represent an important resource for analysis of changes at all levels which emerged in Europe with begin of the industrialization period.

Most digital libraries are made available not only to researchers in a certain Humanities domain (e.g. classical philologists, historians, historical linguists), but also to common users. This fact has posited new requirements to the functionalities offered by the Digital Libraries, and thus imposed the usage of methods from LT for content analysis and content presentation in a form understandable to the end user.

There are several challenges related to the above mentioned issues:

- Lack of adequate training material for real-size applications: although the Digital Libraries usually cover a large number of documents, it is difficult to collect a statistically significant corpus for a period of time in which the language remained unchanged.
- Historical variants of languages lack firmly established syntactic or morphological structures thus the definition of a robust set of rules is very difficult. Historical texts often constitute a mixture of multilingual paragraphs including Latin, Ancient Greek, Slavonic, etc.
- Historical texts contain a large number of anon-standardized abbreviations.
- The conception of the world is somewhat different from ours, which makes it more difficult to build the necessary knowledge bases.

For newspaper collections there are specific questions related to different stages of the whole cycle starting from acquisition of the digital data, conducting the research until publication of the final research results like use of incomplete OCR, available selection of digital newspapers, or copyright limited access to digital newspapers. Other relevant issues include: recommended standards and tools for structural (and, perhaps, lexical) annotation; methods for quality control of every step in the process (i.e., digitisation, transcription, annotation, tools, web services, etc); specific tools for

layout analysis, for enabling access to the inline images/figures, etc.

This workshop brings together researchers working in the interdisciplinary domain of cultural heritage, specialists in natural language and speech processing working with less-resourced languages as well as key players among Linked Open Data initiatives. They are expected to analyse problems and brainstorm solutions in the automatic analysis of historical documents, uni- or multimedia, their deep annotation and interlinking. The workshop builds on successful previous initiatives in this domain at LREC 2010, 2012, and RANLP 2011.

We received a considerable number of papers from which we selected 13, grouped in three sections, centered around following topics:

1. Language Resources for historical documents
2. Historical Newspaper Archives
3. Tools for analysis of historical documents

We are particulary grateful to our invited speaker, Eiríkur Rögnvaldsson, who will present the case of Old Icelandic as historical language and how modern technologies can be used to process it.

We would like to thank all members of the Programme Committee who reviewed in very short time a large number of papers and gave very useful feedback

The workshop is endorsed by the CLARIN Infrastructure Project http://www.clarin.eu

Kristín Bjarnadóttir, Mathew Driscoll, Steven Krauwer, Stelios Piperidis, Cristina Vertan, Martin Wynne

Session Invited Talk

Monday 26 May, 9:30 – 10:30
Chairperson: Steven Krauwer

Old languages, new technologies: The case of Icelandic

Eiríkur Rögnvaldsson

Abstract
In the past few years, interest in developing language technology tools and resources for historical languages or older stages of modern languages has grown considerably, and a number of experiments in adapting existing language technology tools and resources to older variants of the languages in question have been made. The reasons for this increased interest can vary. One is that more and more historical texts are becoming digitized and thus amenable to language technology work. As a result, researchers from many disciplines are starting to realize that they could benefit from being able to search these texts and analyze them with all sorts of language technology tools. Another reason is that since historical texts often exhibit considerable variation in spelling and morphology, they pose great challenges to existing language technology tools and methods developed for modern standardized texts. Thus, many language technology researchers see historical texts as a good test bed for developing and enhancing their methods and tools. In many ways, Icelandic is well suited for being such a test bed. Icelandic has a relatively large corpus of texts from all stages in its recorded history, starting with a number of narrative texts from the 13th and 14th centuries such as the well-known Family Sagas and the so-called Contemporary Sagas. Most importantly, however, Icelandic has changed less during the last thousand years than most or all other languages with a recorded history. True, the sound system has changed radically, especially the vowel system, but these changes are for the most part not reflected in the spelling. There are a number of changes in the syntax, especially as regards word order, but importantly for language technology work, morphological changes are minimal – the Modern Icelandic inflectional system is almost identical to the Old Icelandic system. Of course, the vocabulary has changed considerably, but since the changes are mainly due to new words being added rather than to old words becoming obsolete, these changes do not pose problems for the adaptation of language technology tools to older stages of the language.

Like most other historical languages, older stages of Icelandic show great variation in spelling, even though it may be mentioned that in the 12th century, shortly after Icelanders started writing in Latin letters, an unknown person usually referred to as the First Grammarian made an attempt to standardize Icelandic spelling in a famous essay called the First Grammatical Treatise. This attempt was not successful, and up to the beginning of the 19th century, everyone used his or her own spelling, usually reflecting a mixture of their own pronunciation and the spelling of the manuscripts they had been exposed to, accompanied by considerable intra-scribal variation. With the advent of periodicals around 1800, and especially after the advent of weekly newspapers around 1850, the spelling gradually became more and more uniform and around 1900, a commonly agreed standard had emerged. Since serious work on Icelandic language technology started some 12 years ago, several important resources and tools for Modern Icelandic have been built, such as the Database of Modern Icelandic Inflections, the Tagged Icelandic
Corpus, and the IceNLP package including a POS tagger, a shallow parser and a lemmatizer, to name the most important ones. As a result of the META-NORD project, most of the existing tools and resources are now open and free for everyone to use under standard licenses (GNU and Creative Commons).
In my talk, I will give an overview of the experiments that have been made in adapting and developing language technology tools for older stages of Icelandic. This includes the building of a

parsed historical corpus (IcePaHC) spanning almost ten centuries; adapting POS taggers developed for Modern Icelandic to tagging Old Icelandic texts; developing tools to correct and normalize OCR scanned text of 19th century periodicals and newspapers; and using the IceNLP package in preparing electronic editions of Old Icelandic texts. These experiments have approached similar problems in different ways and I will compare their methods and assess their success in solving the problems.

Session Language Resources

Monday 26 May, 11:00 – 14:40
Chairperson: Kristín Bjarnadóttir

A Personal Name Treebank and Name Parser to Support Searching and Matching of People Names in Historical and Multilingual Contexts
Patrick Schone

Abstract
Personal names are often key elements desired from the processing of historical documents. An understanding of name syntax can be very valuable in aiding automation and analysis. Yet current grammatical parsers classify personal names as merely noun phrases with proper nouns as constituents. To that end, we have created a Personal Name Treebank (PNTB) and associated Statistical Personal Name Parser (SPNP) which are designed to carefully analyze syntactic structure of personal names. The PNTB consists of almost 10 million instances of constituency-parsed personal names attached to genealogically-related contexts. These instances are drawn from almost 200 different countries across millenia. The SPNP leverages the PNTB to achieve 94.4% parse constituency accuracy on a huge held-out set of names. The PNTB and SPNP represent significant new resources which we intend to make available to the research community. To our knowledge, comparable resources have never previously been created.

Corpas na Gaeilge (1882-1926): Integrating Historical and Modern Irish Texts
Elaine Uí Dhonnchadha, Kevin Scannell, Ruairí Ó hUiginn, Eilís Ní Mhearraí, Máire Nic Mhaoláin, Brian Ó Raghallaigh, Gregory Toner, Séamus Mac Mathúna, Déirdre D'Auria, Eithne Ní Ghallchobhair and Niall O'Leary

Abstract
This paper describes the processing of a corpus of seven million words of Irish texts from the period 1882-1926. The texts which have been captured by typing or optical character recognition are processed for the purpose of lexicography. Firstly, all historical and dialectal word forms are annotated with their modern standard equivalents using software developed for this purpose. Then, using the modern standard annotations, the texts are processed using an existing finite-state morphological analyser and part-of-speech tagger. This method enables us to retain the original historical text, and at the same time have full corpus-searching capabilities using modern lemmas and inflected forms (one can also use the historical forms). It also makes use of existing NLP tools for modern Irish, and enables integration of historical and modern Irish corpora.

Language resources for early Modern Icelandic
Ásta Svavarsdóttir, Sigrún Helgadóttir and Guðrún Kvaran

Abstract
This paper describes the compilation of a language corpus of early Modern Icelandic, intended for research in linguistics and lexicography. The texts are extracted from a digital library, accessible on the website Tímarit.is, containing scanned images of individual pages and OCR read text from all Icelandic newspapers and periodicals from that period. In its present form this resource does not fulfill all needs of linguists and lexicographers, due mainly to errors in the digitized texts, the lack of annotation, and limited search possibilities. To create a new language corpus from this text material with the time and money available, methods and tools for automatic or semi-automatic correction of OCR errors had to be developed. The text was to be corrected according to the originals, without any standardization, which poses various challenges in the construction of the corpus. These connect to the correction process itself, the possibilities of using available tools for tagging and lemmatizing, as well as the design of search functions and interface. The solution was to build a parallel corpus with two layers, one with diplomatic text and the other with a standardized modern version of the same text, with mapping between the two.

Named Entities in Court: The MarineLives Corpus
Dominique Ritze, Caecilia Zirn, Colin Greenstreet, Kai Eckert and Simone Paolo Ponzetto

Abstract
In this paper, we introduce the MarineLives corpus. This consists of a collection of anually-transcribed historical records of the English High Court of Admiralty between 1650 and 1669. The transcriptions are obtained collaboratively in an open and transparent process, and are made freely available for the research community. We conduct first experiments with off-the-shelf state-of-the-art Natural Language Processing (NLP) tools to extract named entities from this corpus. In particular, we investigate to what degree the historical language and the highly specific domain of the document affects the results. We find that non-trivial challenges lie ahead and that domain-specific approaches are needed to improve the extraction results.

Building Less Fragmentary Cuneiform Corpora: Challenges and Steps Toward Solutions
Stephen Tyndall

Abstract
This paper examines the particular problems of connecting fragmentary cuneiform material into larger corpora. There are several distinct layers of challenges in this kind of corpus connection project at the orthographic, morphological, and full-text levels of textual analysis. This paper will present these challenges and lay out a pathway by which they may be overcome. Further, specific techniques, particularly the creation of morphological schemata for eventual use as a preprocessing step for classification and clustering, are discussed. These challenges and techniques are presented with respect to their applicability and utility in building these corpora, with the eventual goal of creating fuller, more connected, and more useful cuneiform corpora for scholarly work.

A diachronic corpus of Icelandic poetic texts
Thorhallur Eythorsson, Bjarki Karlsson and Sigríður Sæunn Sigurðardóttir, Greinir skáldskapar:

Abstract
Greinir skáldskapar is a part-of-speech tagged, lemmatized and syntactically annotated corpus of historical Icelandic poetry (http://bragi.info/greinir/), which it is also annotated for phonological and metrical factors. The purpose of the corpus is to enable an integrated search for syntax (dependency

grammar model), phonology and metrics. The database employs a preprogrammed query system, where the user chooses what to search for from an option window. Currently, the options include query possibilities concerning metrics, syllable structure, compound words, grammatical categories, clause structure, syntactic dominance and word search. Different factors can be combined to search for elements that fulfill more than one condition. It is also possible to conduct a search in a specific poem only, or particular poetic genre. A comparison of Greinir skáldskapar with a diachronic corpus of Icelandic prose (IcePaHC) reveals that despite differences they complement each other in a number of ways.

Session Historical Newspaper Archives

Monday 26 May, 14:00 – 16:00
Chairperson: Martin Wyne

On the Role of Historical Newspapers in Disseminating Foreign Words in German

Oliver Pfefferkorn and Peter Fankhauser

Abstract
Newspapers became extremely popular in Germany during the 18th and 19th century, and thus increasingly influential for modern German. However, due to the lack of digitized historical newspaper corpora for German, this influence could not be analyzed systematically. In this paper, we introduce the Mannheim Corpus of Digital Newspapers and Magazines, which in its current release comprises 21 newspapers and magazines from the 18th and 19 century. With over 4.1 Mio tokens in about 650 volumes it currently constitutes the largest historical corpus dedicated to newspapers in German. We briefly discuss the prospect of the corpus for analyzing the evolution of news as a genre in its own right and the influence of contextual parameters such as region and register on the language of news. We then focus on one historically influential aspect of newspapers -- their role in disseminating foreign words in German. Our preliminary quantitative results indeed indicate that newspapers use foreign words significantly more frequently than other genres, in particular belle lettres.

Digital libraries of historical Icelandic newspapers, periodicals, magazines and old printed books

Örn Hrafnkelsson and Jökull Sævarsson

Abstract
The National and University Library of Iceland has from its opening in 1994 taken active part in retroactive digitization of its national collections and made them freely available for its users on the internet. Two projects, digitizing of historical newspapers, magazines and periodicals (timarit.is), and digitizing of historical Icelandic printed books (baekur.is), have the aim to give full access to the material. Both the printed pages and the printed text which opens up new methods for the academic world in its research. The metadata and the text of the material is to be made available for its users.

Historical Newspapers & Journals for the DTA
Susanne Haaf and Matthias Schulz

Abstract
In this paper we present work in progress on the digitization of historical German newspapers in the context of the DFG-funded project Deutsches Textarchiv (DTA) at the Berlin-Brandenburg Academy of Sciences and Humanities. Currently, the DTA core corpus consists of a selection of 1,300 works (ca. 419,000 pages; ca. 97 million tokens) which is balanced with regard to time of creation, genre and theme. It contains mostly monographic works from different disciplines and genres as well as a selection of scientific articles. Though it includes a balanced selection of relevant functional literature, newspaper texts have not been in the scope of text digitization for the DTA, yet. However, the text type newspaper constitutes a significant representative of widespread functional literature and is thus an important source for the usage and development of the German colloquial language. Thus, in the course of the module DTA Extensions (DTAE) we are currently working on the integration of historical newspapers which were digitized by other, external projects. As examples for the integration of historical newspapers into the DTA we here present four different cooperation projects currently maintained by the DTA.

Session Tools for analysis of historical documents
Monday 26 May, 16:30 – 18:00
Chairperson: Matthew Driscoll

The Journal Fjölnir for Everyone: The Post-Processing of Historical OCR Texts
Jón Daðason, Kristín Bjarnadóttir and Kristján Rúnarsson

Abstract
The journal Fjölnir is a much beloved and romanticized 19th century Icelandic journal, published in 1835-1847, which is accessible in digitized form in the digital libraries of the National and University Library of Iceland. In the 19th century, Icelandic spelling was not standardized, and the Fjölnir texts were used for spelling experimentation. The spelling is therefore very varied. In the project described in this paper, the aim was making the text of Fjölnir accessible on the Web, both in the original spelling, and in modern (standardized) spelling, in a version suitable both for scholars and the general public. The modern version serves two purposes. It makes the text more readable for the general public, and it allows the use of NLP tools made for Modern Icelandic. The post-processing of the OCR texts described in this paper was done with the aid of an interactive spellchecker, based on a noisy channel model. The spellchecker achieved a correction accuracy of up to 71.7% when applied on OCR text, and 84.6% when used to normalize the 19th century text to modern spelling.

IceTagging the "Golden Codex". Using language tools developed for Modern Icelandic on a corpus of Old Norse manuscripts.
Ludger Zeevaert

Abstract
Compared to corpora of texts from modern languages that can be based on texts already existing in electronic form or on electronic versions produced mechanically with OCR, the compilation of a corpus of older Icelandic manuscript texts is a rather laborious and time-consuming task because the texts have to be transcribed manually from the manuscripts. Whereas the current state of the art does not allow for a substantial reduction of the workload needed for the transcriptions, the preparation of the corpus for practical tasks offers promising possibilities for automatic or

semiautomatic procedures. The article describes an attempt to use language tools developed for a corpus of Modern Icelandic texts on an Old Icelandic corpus built on manuscript transcriptions in order to enrich the corpus with information useful for linguistic research. POS-tagging of versions of the manuscript texts transferred to Modern Icelandic spelling with taggers developed for Modern Icelandic delivered absolutely satisfactory results, but only after adapting the procedure of manuscript transcription especially to the needs of the utilised language tools. With a reverse approach, i.e. an adaptation of the tools to the demands of medieval Icelandic manuscript corpora, it might be possible to extend their usability in the construction of a multi-purpose corpus and to profit for their further development from work done in more traditional approaches to textual science.

Computational Analysis of Historical Documents: An Application to Italian War Bulletins in World War I and II
Federico Boschetti, Andrea Cimino, Felice Dell'Orletta, Gianluca Lebani, Lucia Passaro, Paolo Picchi, Giulia Venturi, Simonetta Montemagni and Alessandro Lenci

Abstract
World War (WW) I and II represent crucial landmarks in the history on mankind: They have affected the destiny of whole generations and their consequences are still alive throughout Europe. In this paper we present an ongoing project to carry out a computational analysis of Italian war bulletins in WWI and WWII, by applying state-of-the-art tools for NLP and Information Extraction. The annotated texts and extracted information will be explored with a dedicated Web interface, allowing for multidimensional access and exploration of historical events through space and time.

Discovering and Explaining Knowledge in Multilingual Historical Documents
Cristina Vertan, Walther v. Hahn

Abstract
In this paper, we describe the formatting guidelines for ACM SIG Proceedings. The digitization effort during the last years had as result a relative big number of digital version of old manuscripts and printed books. Through this process the preservation of the text is secured but its availability is still restricted to a small number of scientists who understand the language and are familiarized with obsolete geographical names, historical figures etc. Digital repositories offer however for the first time the opportunity that a broad public gets familiarized with this cultural testimony. It is thus of great importance that some effort is invested in enhancing historical texts with explanations and links to modern knowledge bases. Text technology offers methods, which can support this activity. in this paper we present a first attempt for creating an annotation scheme for tagging linguistic, domain specific and general knowledge in historical texts. We exemplify this on a multilingual text written originally in Latin, translated into German and containing a big number of lexical material in Romanian, Ancient Greek and Latin. We explain the necessity of such annotation, describe the challenges and present first results.

CCURL 2014: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era

26 May 2014

ABSTRACTS

Editors:

Laurette Pretorius, Claudia Soria, Paola Baroni

Workshop Programme

09:15-09:30 – Welcome and Introduction

09:30-10:30 – Invited Talk
Steven Moran, *Under-resourced languages data: from collection to application*

10:30-11:00 – Coffee break

11:00-13:00 – Session 1
Chairperson: Joseph Mariani

11:00-11:30 – Oleg Kapanadze, *The Multilingual GRUG Parallel Treebank – Syntactic Annotation for Under-Resourced Languages*

11:30-12:00 – Martin Benjamin, Paula Radetzky, *Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowdsourcing, and Gamification*

12:00-12:30 – Thierry Declerck, Eveline Wandl-Vogt, Karlheinz Mörth, Claudia Resch, *Towards a Unified Approach for Publishing Regional and Historical Language Resources on the Linked Data Framework*

12:30-13:00 – Delphine Bernhard, *Adding Dialectal Lexicalisations to Linked Open Data Resources: the Example of Alsatian*

13:00-15:00 – Lunch break

13:00-15:00 – Poster Session
Chairpersons: Laurette Pretorius and Claudia Soria

Georg Rehm, Hans Uszkoreit, Ido Dagan, Vartkes Goetcherian, Mehmet Ugur Dogan, Coskun Mermer, Tamás Varadi, Sabine Kirchmeier-Andersen, Gerhard Stickel, Meirion Prys Jones, Stefan Oeter, Sigve Gramstad, *An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age"*

István Endrédy, *Hungarian-Somali-English Online Dictionary and Taxonomy*

Chantal Enguehard, Mathieu Mangeot, *Computerization of African Languages-French Dictionaries*

Uwe Quasthoff, Sonja Bosch, Dirk Goldhahn, *Morphological Analysis for Less-Resourced Languages: Maximum Affix Overlap Applied to Zulu*

Edward O. Ombui, Peter W. Wagacha, Wanjiku Ng'ang'a, *InterlinguaPlus Machine Translation Approach for Under-Resourced Languages: Ekegusii & Swahili*

Ronaldo Martins, *UNLarium: a Crowd-Sourcing Environment for Multilingual Resources*

Anuschka van ´t Hooft, José Luis González Compeán, *Collaborative Language Documentation: the Construction of the Huastec Corpus*

Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, Francis M. Tyers, *Open-Source Infrastructures for Collaborative Work on Under-Resourced Languages*

15:00-16:00 – Session 2
Chairperson: Eveline Wandl-Vogt

15:00-15:30 – Riccardo Del Gratta, Francesca Frontini, Anas Fahad Khan, Joseph Mariani, Claudia Soria, *The LREMap for Under-Resourced Languages*

15:30-16:00 – Dorothee Beermann, Peter Bouda, *Using GrAF for Advanced Convertibility of IGT data*

16:00-16:30 – Coffee break

16:30-17:30 – Session 3
Chairperson: Thierry Declerck

16:30-17:00 – Stefan Daniel Dumitrescu, Tiberiu Boroș, Radu Ion, *Crowd-Sourced, Automatic Speech-Corpora Collection – Building the Romanian Anonymous Speech Corpus*

17:00-17:30 – Katia Kermanidis, Manolis Maragoudakis, Spyros Vosinakis, *Crowdsourcing for the Development of a Hierarchical Ontology in Modern Greek for Creative Advertising*

17:30-18:15 – Discussion

18:15-18:30 – Wrap-up and goodbye

Workshop Organizing Committee

Laurette Pretorius	University of South Africa, South Africa
Claudia Soria	CNR-ILC, Italy
Eveline Wandl-Vogt	Austrian Academy of Sciences, ICLTT, Austria
Thierry Declerck	DFKI GmbH, Language Technology Lab, Germany
Kevin Scannell	St. Louis University, USA
Joseph Mariani	LIMSI-CNRS & IMMI, France

Workshop Programme Committee

Deborah W. Anderson	University of Berkeley, Linguistics, USA
Sabine Bartsch	Technische Universität Darmstadt, Germany
Delphine Bernhard	LILPA, Strasbourg University, France
Bruce Birch	The Minjilang Endangered Languages Publications Project, Australia
Paul Buitelaar	DERI, Galway, Ireland
Peter Bouda	CIDLeS - Interdisciplinary Centre for Social and Language Documentation, Portugal
Steve Cassidy	Macquarie University, Australia
Thierry Declerck	DFKI GmbH, Language Technology Lab, Germany
Vera Ferreira	CIDLeS - Interdisciplinary Centre for Social and Language Documentation, Portugal
Claudia Garad	wikimedia.AT, Austria
Dafydd Gibbon	Bielefeld University, Germany
Oddrun Grønvik	Instituut for lingvistike og nordiske studier, University of Oslo, Norway
Yoshihiko Hayashi	University of Osaka, Japan
Daniel Kaufman	Endangered Language Alliance, USA
Andras Kornai	Hungarian Academy of Sciences, Hungary
Simon Krek	Jožef Stefan Institute, Slovenia
Tobias Kuhn	ETH, Zurich, Switzerland
Joseph Mariani	LIMSI-CNRS & IMMI, France
Steven Moran	Universität Zürich, Switzerland
Kellen Parker	National Tsing Hua University, China
Patrick Paroubek	LIMSI-CNRS, France
Maria Pilar Perea i Sabater	Universitat de Barcelona, Spain
Laurette Pretorius	University of South Africa, South Africa
Leonel Ruiz Miyares	Centro de Linguistica Aplicada (CLA), Cuba
Kevin Scannell	St. Louis University, USA
Ulrich Schäfer	DFKI GmbH, Germany
Claudia Soria	CNR-ILC, Italy
Nick Thieberger	University of Melbourne, Australia
Michael Zock	LIF-CNRS, France

Preface

The LREC 2014 Workshop on "Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era" (CCURL 2014) has its origin in the imperative of cultural and language diversity and in the basic right of all communities, all languages and all cultures to be "first class citizens" in an age driven by information, knowledge and understanding. In this spirit, the focus of this first CCURL Workshop is on two strategic approaches by which under-resourced languages can elevate themselves to levels of development that are potentially comparable to well-resourced, technologically advanced languages, viz. using the crowd and collaborative platforms, and using technologies of interoperability with well-developed languages and Linked Data.

Specific questions that the Workshop addresses include the following:

- How can collaborative approaches and technologies be fruitfully applied to the development and sharing of resources for under-resourced languages?
- How can small language resources be re-used efficiently and effectively, reach larger audiences and be integrated into applications?
- How can they be stored, exposed and accessed by end users and applications?
- How can research on such languages benefit from semantic and semantic web technologies, and specifically the Linked Data framework?

All the papers accepted for the Workshop address at least one of these questions, thereby making a noteworthy contribution to the relevant scholarly literature and to the technological development of a wide variety of under-resourced languages.

Each of the sixteen accepted papers was reviewed by at least three members of the Programme Committee, eight of which are presented as oral presentations and eight as posters.

We look forward to collaboratively and computationally building on this new tradition of CCURL in the future for the continued benefit of all the under-resourced languages of the world!

The Workshop Organizers

Laurette Pretorius – University of South Africa, South Africa
Claudia Soria – CNR-ILC, Italy
Eveline Wandl-Vogt – Austrian Academy of Sciences, ICLTT, Austria
Thierry Declerck – DFKI GmbH, Language Technology Lab, Germany
Kevin Scannell – St. Louis University, USA
Joseph Mariani – LIMSI-CNRS & IMMI, France

Session 1

Monday 26 May, 11:00-13:00
Chairperson: Joseph Mariani

The Multilingual GRUG Parallel Treebank – Syntactic Annotation for Under-Resourced Languages

Oleg Kapanadze

In this paper, we describe outcomes of an undertaking on building Treebanks for underresourced languages Georgian, Russian, Ukrainian, and German - one of the "major" languages in the NLT world (Hence, the treebank 's name – GRUG). The monolingual parallel sentences in four languages were syntactically annotated manually using the Synpathy tool. The tagsets follow an adapted version of the German TIGER guidelines with necessary changes relevant for the Georgian, the Russian and the Ukrainian languages grammar formal description. An output of the monolingual syntactic annotation is in the TIGER-XML format. Alignment of monolingual repository into the bilingual Treebanks was done by the Stockholm TreeAligner software. The parallel treebank resources developed in the GRUG project can be viewed at the URL of Saarland and Bergen Universities: http://fedora.clarin-d.uni-saarland.de/grug/ , http://clarino.uib.no/iness.

Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowdsourcing, and Gamification

Martin Benjamin, Paula Radetzky

This paper looks at the challenges that the Kamusi Project faces for acquiring open lexical data for less-resourced languages (LRLs), of a range, depth, and quality that can be useful within Human Language Technology (HLT). These challenges include accessing and reforming existing lexicons into interoperable data, recruiting language specialists and citizen linguists, and obtaining large volumes of quality input from the crowd. We introduce our crowdsourcing model, specifically (1) motivating participation using a "play to pay" system, games, social rewards, and material prizes; (2) steering the crowd to contribute structured and reliable data via targeted questions; and (3) evaluating participants' input through crowd validation and statistical analysis to ensure that only trust-worthy material is incorporated into Kamusi's master database. We discuss the mobile application Kamusi has developed for crowd participation that elicits high-quality structured data directly from each language's speakers through narrow questions that can be answered with a minimum of time and effort. Through the integration of existing lexicons, expert input, and innovative methods of acquiring knowledge from the crowd, an accurate and reliable multilingual dictionary with a focus on LRLs will grow and become available as a free public resource.

Towards a Unified Approach for Publishing Regional and Historical Language Resources on the Linked Data Framework

Thierry Declerck, Eveline Wandl-Vogt, Karlheinz Mörth, Claudia Resch

We describe actual work on porting dialectal dictionaries and historical lexical resources developed at the Austrian Academy of Sciences onto representation languages that are supporting their publication in the Linked (Open) Data framework. We are aiming at a unified representation model that is flexible enough for describing those distinct types of lexical information. The goal is not only to be able to cross-link those resources, but also to link them in the Linked Data cloud with available data sets for highly-resourced languages and to elevate this way the dialectal and historical lexical resources to the same "digital dignity" as the mainstream languages have already gained.

Adding Dialectal Lexicalisations to Linked Open Data Resources: the Example of Alsatian

Delphine Bernhard

This article presents a method to align bilingual lexicons in a resource-poor dialect, namely Alsatian. One issue with Alsatian is that there is no standard and widely-acknowledged spelling convention and a lexeme may therefore have several different written variants. Our proposed method makes use of the double metaphone algorithm adapted to Alsatian in order to bridge the gap between different spellings. Once variant citation forms of the same lexeme have been aligned, they are mapped to BabelNet, a multilingual semantic network (Navigli and Ponzetto, 2012). The mapping relies on the French translations and on cognates for Alsatian words in the English and German languages.

Poster Session

Monday 26 May, 13:00-15:00
Chairpersons: Laurette Pretorius and Claudia Soria

An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age"

Georg Rehm, Hans Uszkoreit, Ido Dagan, Vartkes Goetcherian, Mehmet Ugur Dogan, Coskun Mermer, Tamás Varadi, Sabine Kirchmeier-Andersen, Gerhard Stickel, Meirion Prys Jones, Stefan Oeter, Sigve Gramstad

This paper extends and updates the cross-language comparison of LT support for 30 European languages as published in the META-NET Language White Paper Series. The updated comparison confirms the original results and paints an alarming picture: it demonstrates that there are even more dramatic differences in LT support between the European languages.

Hungarian-Somali-English Online Dictionary and Taxonomy

István Endrédy

Background. The number of Somalis coming to Europe has increased substantially in recent years. Most of them do not speak any foreign language, only Somali, but a few of them speak English as well. Aims. A simple and useful online dictionary would help Somalis in everyday life. It should be online (with easy access from anywhere) and it has to handle billions of word forms, as Hungarian is heavily agglutinative. It should handle typos as the users are not advanced speakers of the foreign languages of the dictionary. It should pronounce words, as these languages have different phonetic sets. It should be fast with good precision because users do not like to wait. And last but not least, it should support an overview of the vocabulary of a given topic. Method. A vocabulary (2000 entries) and a taxonomy (200 nodes) was created by a team (an editor and a native Somali speaker) in an Excel table. This content was converted into a relational database (mysql), and it got an online user interface based on php and jqueryui. Stemmer and text-to-speech modules were included and implemented as a web service. Typos were handled with query extension. Results. Although the dictionary lookup process does stemming with a web service and makes a query extension process, it is very fast (100-300ms per query). It can pronounce every Hungarian word and expression owing to the text-to-speech web service. Conclusion. This dictionary was opened to the public in October, 2013. (http://qaamuus.rmk.hu/en) The next step is the creation of a user interface optimised for mobile devices.

Computerization of African Languages-French Dictionaries

Chantal Enguehard, Mathieu Mangeot

This paper relates work done during the DiLAF project. It consists in converting 5 bilingual African language-French dictionaries originally in Word format into XML following the LMF model. The languages processed are Bambara, Hausa, Kanuri, Tamajaq and Songhai-zarma, still considered as under-resourced languages concerning Natural Language Processing tools. Once converted, the

dictionaries are available online on the Jibiki platform for lookup and modification. The DiLAF project is first presented. A description of each dictionary follows. Then, the conversion methodology from .doc format to XML files is presented. A specific point on the usage of Unicode follows. Then, each step of the conversion into XML and LMF is detailed. The last part presents the Jibiki lexical resources management platform used for the project.

Morphological Analysis for Less-Resourced Languages: Maximum Affix Overlap Applied to Zulu

Uwe Quasthoff, Sonja Bosch, Dirk Goldhahn

The paper describes a collaboration approach in progress for morphological analysis of less-resourced languages. The approach is based on firstly, a language-independent machine learning algorithm, Maximum Affix Overlap, that generates candidates for morphological decompositions from an initial set of language-specific training data; and secondly, language-dependent post-processing using language specific patterns. In this paper, the Maximum Affix Overlap algorithm is applied to Zulu, a morphologically complex Bantu language. It can be assumed that the algorithm will work for other Bantu languages and possibly other language families as well. With limited training data and a ranking adapted to the language family, the effort for manual verification can be strongly reduced. The machine generated list is manually verified by humans via a web frontend.

InterlinguaPlus Machine Translation Approach for Under-Resourced Languages: Ekegusii & Swahili

Edward O. Ombui, Peter W. Wagacha, Wanjiku Ng'ang'a

This paper elucidates the InterlinguaPlus design and its application in bi-directional text translations between Ekegusii and Kiswahili languages unlike the traditional translation pairs, one-by-one. Therefore, any of the languages can be the source or target language. The first section is an overview of the project, which is followed by a brief review of Machine Translation. The next section discusses the implementation of the system using Carabao's open machine translation framework and the results obtained. So far, the translation results have been plausible particularly for the resource-scarce local languages and clearly affirm morphological similarities inherent in Bantu languages.

UNLarium: a Crowd-Sourcing Environment for Multilingual Resources

Ronaldo Martins

We present the UNLarium, a web-based integrated development environment for creating, editing, validating, storing, normalising and exchanging language resources for multilingual natural language processing. Conceived for the UNL Lexical Framework, the UNLarium provides semantic accessibility to language constrained data, as it interconnects lexical units from several different languages, through taxonomic and non-taxonomic relations, representing not only necessary but also typical associations, obtained from machine learning and human input, in order to create an incremental and dynamic map of the human knowledge.

Collaborative Language Documentation: the Construction of the Huastec Corpus

Anuschka van ´t Hooft, José Luis González Compeán

In this paper, we describe the design and functioning of a web-based platform called Nenek, which aims to be an on-going language documentation project for the Huastec language. In Nenek, speakers, linguistic associations, government instances and researchers work together to construct a centralized repository of materials about the Huastec language. Nenek not only organizes different types of contents in repositories, it also uses this information to create online tools such as a searchable database with documents on Huastec language and culture, E-dictionaries and spell checkers. Nenek is also a monolingual social network in which users discuss contents on the

platform. Until now, the speakers have created a monolingual E-dictionary and we have initiated an on-going process of the construction of a repository of written texts in the Huastec language. In this context, we have been able to localize and digitally archive documents in other formats (audios, videos, images), yet the retrieval, creation, storage, and documentation of this type of materials is still in a preliminary phase. In this presentation, we want to present the general methodology of the project.

Open-Source Infrastructures for Collaborative Work on Under-Resourced Languages

Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, Francis M. Tyers

In order to support crowd sourcing for a language, certain social and technical prerequisites must be met. Both the size of the community and the level of technical support available are important factors. Many language communities are too small to be able to support a crowd-sourcing approach to building language-technology resources, while others have a large enough community but require a platform that relieves the need to develop all the technical and computational-linguistic know how needed to actually run a project successfully. This article covers the languages being worked on in the Giellatekno/Divvun and Apertium infrastructures. Giellatekno is a language-technology research group, Divvun is a product development group and both work primarily on the Sámi languages. Apertium is a free/open-source project primarily working on machine translation. We use Wikipedia as an indicator to divide the set of languages that we work on into two groups: those that can support traditional crowdsourcing, and those that do not. We find that the languages being worked on in the Giellatekno/Divvun infrastructure largely fall into the latter group, while the languages in the Apertium infrastructure fall mostly into the former group. Regardless of the ability of a language community to support traditional crowdsourcing, there is in all cases the necessity to provide a technical infrastructure to back up any linguistic work. We present two infrastructures, the Giellatekno/Divvun infrastructure and the Apertium infrastructure and show that while both groups of language communities would not be able to develop language technology on their own, using the infrastructures that we present they have been quite successful.

Session 2

Monday 26 May, 15:00-16:00
Chairperson: Eveline Wandl-Vogt

The LREMap for Under-Resourced Languages

Riccardo Del Gratta, Francesca Frontini, Anas Fahad Khan, Joseph Mariani, Claudia Soria

A complete picture of currently available language resources and technologies for the under-resourced languages of Europe is still lacking. Yet this would help policy makers, researchers and developers enormously in planning a roadmap for providing all languages with the necessary instruments to act as fully equipped languages in the digital era. In this paper we introduce the LRE Map and show its utility for documenting available language resources and technologies for under-resourced languages. The importance of the serialization of the LREMap into (L)LOD along with the possibility of its connection to a wider world is also introduced.

Using GrAF for Advanced Convertibility of IGT data

Dorothee Beermann, Peter Bouda

With the growing availability of multi-lingual, multi-media and multi-layered corpora also for lesser-documented languages, and with a growing number of tools that allow their exploitation, working with corpora has attracted the interest of researchers from the theoretical as well as the applied fields. But always when information from different sources is combined, the pertaining lack of interoperability represents a problem. This is in particular a challenge for corpora from endangered and lesser described languages since they often originate from work by individual researchers and

small projects, using different methodologies and different tools. Before this material can become a true resource, well-known differences in the physical as well as the conceptual data structure must be leveraged against ways of future data use and exploitation. Working with Interlinear Glossed Text (IGT), which is a common annotation format for linguistic data from lesser described languages, we will use GrAF to achieve Advanced Convertibility. Our goal is to build data bridges between a number of linguistic tools. As a result data will become mobile across applications.

Session 3

Monday 26 May, 16:30-17:30
Chairperson: Thierry Declerck

Crowd-Sourced, Automatic Speech-Corpora Collection – Building the Romanian Anonymous Speech Corpus

Stefan Daniel Dumitrescu, Tiberiu Boroș, Radu Ion

Taking the example of other successful initiatives such as VoxForge, we applied the concept of crowd-sourcing to respond to a particular need: the lack of free-speech, time-aligned, multi-user corpora for the Romanian language. Such speech corpora represent a valuable asset for spoken language processing application because they offer the means to (1) train and test acoustic models and (2) develop and validate various methods and techniques that are intended to enhance today's ASR and TTS technologies.

Crowdsourcing for the Development of a Hierarchical Ontology in Modern Greek for Creative Advertising

Katia Kermanidis, Manolis Maragoudakis, Spyros Vosinakis

This paper describes the collaborative development of a hierarchical ontology in the domain of television advertisement. The language addressed is Modern Greek, one of the not widely spoken and not-richly-equipped-with-resources languages. The population of the ontology is achieved through collaborative crowdsourcing, i.e. players annotate ad video content through a multi-player videogame, implemented especially for this purpose. The provided annotations concern the ad content, its production values, its impact, and they constitute the ontology terms and concepts. Dependencies, correlations, statistical information and knowledge governing the ontology terms and concepts are to be revealed through data mining and machine learning techniques. The extracted knowledge constitutes the core of a support tool, i.e. a semantic thesaurus, which will help ad designers in the brainstorming process of creating a new ad campaign. Unlike existing creativity support models, that are static and depend on expert knowledge, thereby hurting creativity, the proposed support tool is generic in nature (as it is based on a collaborative crowdsourcing-based semantic thesaurus), dynamic and minimally restricting the brainstorming process.

Come Hack with OpeNER!

26 May 2014

ABSTRACTS

Editors:

Seán Gaines, Vicomtech-IK4

Montse Cuadros, Vicomtech-IK4

Workshop Programme

26 May 2014

9:00 – 9:20 **Introduction by Workshop Chair**

9:20 – 10:30 **Tutorial: OpeNER technology**

10:30 – 11:00 **Coffee break**

11:00 –11:45 **Demo/Posters**

Carlo Aliprandi, Sara Pupi and Giulia di Pietro, *Ent-it-UP: a Sentiment Analysis system based on OpeNER cloud services*

Jordi Atserias, Marieke van Erp, Isa Maks, German Rigau and J. Fernando Sánchez-Rada, *EuroLoveMap: Confronting feelings from News*

Estela Saquete and Sonia Vázquez, *Improving reading comprehension for hearing impared students using Natural Language Processing*

Aitor García Pablos, Montse Cuadros, Seán Gaines and German Rigau, *OpeNER demo: Open Polarity Enhanced Named Entity Recognition*

Andoni Azpeitia, Alexandra Balahur, Montse Cuadros, Antske Fokkens and Ruben Izquierdo Bevia, *The Snowball effect: following opinions on controversial topics*

Stefano Cresci, Andrea D'Errico, Davide Gazzé, Angelica Lo Duca, Andrea Marchetti and Maurizio Tesconi, *Tour-pedia: a Web Application for Sentiment Visualization in Tourism Domain*

12:00 –13:00 **Lunch break**

12:00 – 16:00 **Hackathon**

16:00 – 16:30 **Coffee break**

16:30 – 17:30 **Results presentation**

Workshop Organizers

Rodrigo Agerri	EHU/UPV
Montse Cuadros	Vicomtech-IK4
Francesca Frontini	CNR-ILC
Seán Gaines	Vicomtech-IK4
Ruben Izquierdo	VUA
Wilco van Duinkerken	Olery

Workshop Programme Committee

Carlo Aliprandi	Synthema
Andoni Azpeitia	Vicomtech-IK4
Aitor Garcia-Pablos	Vicomtech-IK4
Angelica Lo Duca	CNR-IIT
Isa Maks	VUA
Andrea Marchetti	CNR-IIT
Monica Monachini	CNR-ILC
German Rigau	EHU/UPV
Piek Vossen	VUA

Introduction

The OpeNER team is delighted to present a **Tutorial** and a **Hackathon** together in a one-day **workshop** on multilingual Sentiment Analysis and Named Entity Resolution using the OpeNER NLP pipelines as web services in the Cloud.

OpeNER hopes to repeat the success from the July 2013 Amsterdam Hackathon (http://www.opener-project.org/2013/07/18/opener-hackathon-in-amsterdam/) in which a broad spectrum of real end user SMEs, Micro-SMEs, Freelancers and even a few from technology giants, built creative applications using the OpeNER webservices. For examples of the applications built follow the URL provided above.

The proposed workshop will present briefly the project, and all the technology (http://opener-project.github.io/) multilingual NLP tools and resources created within the project. Additionally, it will be a slot for presentations of demos created before the Hackathon and presented in the call for papers.

The workshop will be complemented by a half day Hackathon. The Hackathon will encourage participants to form ad hoc multidisciplinary teams, brainstorm an idea, implement it and present a demo from which a winner will be picked by popular vote. Most of the "core developers" of the OpeNER pipeline technology will be available to help you out and get started.

All participants will be given access to the collateral needed such as NLP tools and resources in six languages beforehand from publicly deployed web services. As of writing the initial versions of the services are publically available at http://opener.olery.com. In order to present a demo or paper to the workshop the only thing that needs to be added is imagination.

Session Demo/Posters

Monday 26 May 11:00- 11:45
Chairperson: Seán Gaines

Ent-it-UP: a Sentiment Analysis system based on OpeNER cloud services
Carlo Aliprandi, Sara Pupi and Giulia di Pietro

In this paper we present a web application that exploit OpeNER cloud services. Ent-it-UP monitors Social Media and traditional Mass Media contents, performing multilingual Named Entity Recognition and Sentiment Analysis. The goal of Ent-it-Up is to carry out statistics about retrieved entities and display results in a user friendly interface. In this way the final consumer can easily have an insight about what people think of a certain product, brand or, in general, about a certain topic.

EuroLoveMap: Confronting feelings from News
Jordi Atserias, Marieke van Erp, Isa Maks, German Rigau and J. Fernando Sánchez-Rada

Opinion mining is a natural language analysis task aimed at obtaining the overall sentiment regarding a particular topic. This paper presents a very preliminary prototype developed in just one day during the hackathon organized by the OpeNER project in Amsterdam last year. Our team called Napoleon used the OpeNER infrastructure developed during the first year of the project, to process a large set of news articles in six different languages. Using these tools an overall sentiment analysis was obtained for each news article. Then, for each language, we gathered all news mentioning a particular topic. We show how the results can be easily confronted on an interactive map.

Improving reading comprehension for hearing impared students using Natural Language Processing
Estela Saquete and Sonia Vázquez

The main objective of this paper is developing a tool capable of transforming educational texts in Spanish into very easy reading texts by using Natural Language Processing (NLP). This tool is mainly oriented to support people with problems in reading comprehension, for instance, deaf people or people who are learning a new language. The process of simplification and enrichment of texts consists of the automatically detection of linguistic features of input texts and: a) the reduction and removal of obstacles, but preserving in all cases the original meaning of the text, and b) the enrichment of texts using different open tools.

OpeNER demo: Open Polarity Enhanced Named Entity Recognition
Aitor García Pablos, Montse Cuadros, Seán Gaines and German Rigau

OpeNER is a project funded by the European Commission under the 7th Framework Programme. Its acronym means Open Polarity Enhanced Named Entity Recognition. OpeNER main goal is to provide a set of open and ready to use tools to perform some NLP tasks in six languages including English, Spanish, Italian, Dutch, German and French. In order to display these OpeNER analysis output in a format suitable for a non-expert human reader we have developed a Web application to display this content in different ways. This Web application should serve as a demonstration of some of the OpeNER modules capabilities.

The Snowball effect: following opinions on controversial topics
Andoni Azpeitia, Alexandra Balahur, Montse Cuadros, Antske Fokkens and Ruben Izquierdo Bevia,

This paper describes a practical application of the OpeNER project's technology to find trending topics in different media sources and in different languages. In this case, we have followed the global scandal on leaking data, Snowden. We applied a rule-based opinion mining tool and the results show a diversity of opinions depending on the language in terms of the language the sources were analyzed and also a interesting opinion division between people pro-Snowden and pro-US.

Tour-pedia: a Web Application for Sentiment Visualization in Tourism Domain
Stefano Cresci, Andrea D'Errico, Davide Gazzé, Angelica Lo Duca, Andrea Marchetti and Maurizio Tesconi

Tour-pedia is a Web application which shows users' sentiments of touristic locations of some of the most important cities and regions in Europe. It is implemented within the OpeNER project [1], which aims to provide a pipeline for processing natural language. More specifically, Tour-pedia, exploits the OpeNER pipeline to analyse users' reviews on places. All reviews are extracted from social media. Once analysed, each review is associated to a rate, which ranges from 0 to 100. The sentiment of each place is calculated as a function of all the sentiments of reviews on that place. As a result, Tour-pedia shows all the places and their users' sentiments on a map

10th Joint ACL – ISO Workshop on Interoperable Semantic Annotation

ISA-10

26 May 2014

ABSTRACTS

Harry Bunt, Editor

Workshop Programme

08.30 – 08:50 On-site registration
08:50 -- 09:00 Opening by Workshop Chair

09:00 -- 10:30 Session A
09:00 -- 09:30 Hans-Ulrich Krieger, *A Detailed Comparison of Seven Approaches for the Annotation of Time-Dependent Factual Knowledge in RDF and OWL*

09:30 -- 10:00 Antske Fokkens, Aitor Soroa, Zuhaitz Beloki, Niels Ockeloen, Piek Vossen, German Rigau and Willem-Robert van Hage, *NAF and GAF: Linking Linguistic Annotations*

10:00 -- 10:15 Johan Bos, *Semantic Annotation Issues in Parallel Meaning Banking*

10:15 --10:30 Assaf Toledo, Stavroula Alexandropoulou, Sophie Chesney, Robert Grimm, Pepijn Kokke, Benno Kruit, Kyriaki Neophytou, Antony Nguyen and Yoad Winter, *A Proof-Based Annotation Platform of Textual Entailment*

10:30 – 11:00 Coffee break

11:00 -- 13:00 Session B
11:00 -- 11:15 Bolette Pedersen, Sanni Nimb, Sussi Olsen, Anders Soegaard and Nnicola Soerensen, *Semantic Annotation of the Danish CLARIN Reference Corpus*
11:15 -- 11:45 Kiyong Lee, *Semantic Annotation of Anaphoric Links in Language*
11:45 -- 12:00 Laurette Pretorius and Sonja Bosch, *Towards extending the ISOcat Data Category Registry with Zulu Morphosyntax*
12:00 -- 13:00 Harry Bunt, Kiyong Lee, Martha Palmer, Rashmi Prasad, James Pustejovsky and Annie Zaenen, *ISO Projects on the development of international standards for the annotation of various types of semantic information*

13:00 – 14:00 Lunch break

14:00 -- 16:00 Session C
14:00 -- 14:30 Volha Petukhova, *Understanding Questions and Finding Answers: Semantic Relation Annotation to Compute the Expected Answer Type*
14:30 -- 14:45 Susan Windisch Brown, *From Visual Prototypes of Action to Metaphors: Extending the IMAGACT Ontology of Action to Secondary Meanings*

14:45 -- 15:15 Ekaterina Lapshinova-Koltunski and Kerstin Anna Kunz, *Annotating Cohesion for Multillingual Analysis*
15:15 -- 16:00 Poster session: elevator pitches followed by poster visits
Leon Derczynski and Kalina Bontcheva: *Spatio-Temporal Grounding of Claims Made on the Web in Pheme*
Mathieu Roche: *How to Exploit Paralinguistic Features to Identify Acronyms in Texts*
Sungho Shin, Hanmin Jung, Inga Hannemann and Mun Yong Yi: *Lessons Learned from Manual Evaluation of NER Results by Domain Experts*
Milan Tofiloski, Fred Popowich and Evan Zhang: *Annotating Discourse Zones in Medical Encounters*
Yu Jie Seah and Francis Bond: *Annotation of Pronouns in a Multilingual Corpus of*

Mandarin Chinese, English and Japanese

16:00 --16:30 Coffee break

16:30 -- 18:00 Session D

16:30 -- 17:00 Elisabetta Jezek, Laure Vieu, Fabio Massimo Zanzotto, Guido Vetere, Alessandro Oltramari, Aldo Gangemi and Rossella Vanvara, *Extending `Senso Comune' with Semantic Role Sets*

17:00 -- 17:30 Paulo Quaresma, Amália Mendes, Iris Hendrickx and Teresa Gonçalves *Automatic Tagging of Modality: Identifying Triggers and Modal Values*

17:30 -- 18:00 Rui Correia, Nuno Mamede, Jorge Baptista and Maxina Eskenazi, *Using the Crowd to Annotate Metadiscursive Acts*

18:00 Workshop Closing

Workshop Organizers/Organizing Committee

Harry Bunt	Tilburg University
Nancy Ide	Vassar College, Poughkeepsie, NY
Kiyong Lee	Korea University, Seoul
James Pustejovsky	Brandeis University, Waltham, MA
Laurent Romary	INRIA/Humboldt Universität Berlin

Workshop Programme Committee

Jan Alexandersson	DFKI, Saarbrücken
Paul Buitelaar	National University of Ireland, Galway
Harry Bunt	Tilburg University
Thierry Declerck	DFKI, Saarbrücken
Liesbeth Degand	Université Catholique de Louvain
Alex Chengyu Fang	City University Hong Kong
Anette Frank	Universität Heidelberg
Robert Gaizauskas	University of Sheffield
Koiti Hasida	Tokyo University
Nancy Ide	Vassar College
Elisabetta Jezek	Università degli Studi di Pavia
Michael Kipp	University of Applied Sciences, Augsburg
Inderjeet Mani	Yahoo, Sunnyvale
Martha Palmer	University of Colorado, Boulder
Volha Petukhova	Universität des Saarlandes, Saarbrücken
Andrei Popescu-Belis	Idiap, Martigny, Switzerland
Rarhmi Prasad	University of Wisconsin, Milwaukee
James Pustejovsky	Brandeis University
Laurent Romary	INRIA/Humboldt Universität Berlin
Ted Sanders	Universiteit Utrecht
Thorsten Trippel	University of Bielefeld
Zdenka Uresova	Charles University, Prague
Piek Vossen	Vrije Universiteit Amsterdam
Annie Zaenen	Stanford University

A Detailed Comparison of Seven Approaches for the Annotation of Time-Dependent Factual Knowledge in RDF and OWL

Hans-Ulrich Krieger

Representing time-dependent information has in recent times become increasingly important. Extending OWL relation instances or RDF triples with further temporal arguments is usually realized through the introduction of new individuals that hide the range of arguments of the extended relations. As a result, reasoning and querying with such representations is extremely complex, expensive, and error-prone. In this paper we discuss several well-known approaches to this problem and present their pros and cons. Three of them are compared in more detail, both on a theoretical and on a practical level. We also present schemata for translating triple-based encodings into general tuples, and vice-versa. Concerning query time, our preliminary measurements have shown that a general tuple-based approach can easily outperform triple-based encodings by several orders of magnitude.

NAF and GAF: Linking Linguistic Annotations

Antske Fokkens, Aitor Soroa, Zuhaitz Beloki, Niels Ockeloen, Piek Vossen, German Rigau and Willem Robert van Hage

Interdisciplinary research between computational linguistics and the Semantic Web is increasing. The NLP community makes more and more use of information presented as Linked Data. At the same time, an increasing interest in representing information from text as Linked Data can be observed in the Semantic Web community. This paper presents the representations we use in two projects that involve both directions of interaction between NLP and the Semantic Web. Previous work has show how instances represented in RDF can be linked to text and linguistic annotations using GAF. In this paper, we address how we can make further use of Linked Data by using its principles in linguistic annotations.

Semantic Annotation Issues in Parallel Meaning Banking
Johan Bos

If we try to align meaning representations of translated sentences, we are faced with the following problem: even though concepts and relations ought to be independent from specific natural languages, the non-logical symbols present in meaning representations usually resemble language-specific words. In faithful translations, such symbols can be easily aligned. In informative translations (where more information is provided by the target translation), symbols can be aligned by a symbol denoting an inclusion relation. In loose translations, we need a third combinator to combine symbols with similar but not identical meanings. We show how this can be done with several concrete, nontrivial English-German translation pairs. The resulting formalism is a first step towards constructing parallel meaning banks.

A Proof-Based Annotation Platform of Textual Entailment

Assaf Toledo, Stavroula Alexandropoulou, Sophie Chesney, Robert Grimm, Pepijn Kokke, Benno Kruit, Kyriaki Neophytou, Antony Nguyen and Yoad Winter

We introduce a new platform for annotating inferential phenomena in entailment data, buttressed by a formal semantic model and a proof-system that provide immediate verification of the coherency and completeness of the marked annotations. By integrating a web-based user interface, a formal lexicon, a lambda-calculus engine and an off-the-shelf theorem prover, the platform allows human annotators to mark linguistic phenomena in entailment data (pairs made up of a premise and a hypothesis) and to receive immediate feedback whether their annotations are substantiated: for positive entailment pairs, the system searches for a formal logical proof that the hypothesis follows from the premise; for negative pairs, the system verifies that a counter-model can be constructed. This novel approach facilitates the creation of textual entailment corpora with annotations that are sufficiently coherent and complete for recognizing the entailment relation or lack thereof. A corpus of several hundred annotated entailments is currently being compiled based on the platform and will be available for the research community in the foreseeable future.

Semantic Annotation of the Danish CLARIN Reference Corpus

Bolette Pedersen, Sanni Nimb, Sussi Olsen, Anders Søgaard and Nicolai Sørensen

The newly initiated project "Semantic Processing across Domains" is granted by the Danish Research Council for Culture and Communication and runs for the period 2013-2016. It focuses on Danish as a low-resourced language and aims at increasing the level of technological resources available for the Danish HLT community. A primary project goal is to provide semantically annotated text corpora of Danish following agreed standards and to let these serve as training data for advanced machine learning algorithms which will address data scarcity and domain adaptation as central problem areas. The Danish CLARIN Reference Corpus - supplemented by a selection of additional text types from social media and the web - are being sense and role annotated. We experiment with an adaptation of PropBank roles to Danish as well as with a scalable sense inventory of Danish. This inventory spans from supersense annotations (semantic classes) to wordnet-derived sense annotations which rely on a distinction between ontological types and main and subsenses. The annotation tool WebAnno, which is being developed as part of the German CLARIN project, is applied for the annotation task.

Semantic Annotation of Anaphoric Links in Language

Kiyong Lee

This paper attempts to integrate several existing coreference annotation schemes into an extended annotation scheme AS_{ana}. The proposed AS_{ana} allows some other types of the anaphor-antecedent relation, called 'anaphoric link', than the canonical type of coreference that implies the referential identity between an anaphor and its antecedent. The structure of AS_{ana} itself is very simple, consisting of a single entity type for mentions and a single anaphoric relation, each of which is characterized by a small set of attribute-value specifications. Constrained by these specifications, AS_{ana} supports a two-step annotation procedure: For a given text T, (1) identification of a set of mentions M in the text T that refer to something in the universe of discourse referents as its markables, (2a) identification of a set of pairs of the mentions in M that are anaphorically related, and (2b) specification of the type of such a relation.

Towards Extending the ISOcat Data Category Registry with Zulu Morphosyntax

Laurette Pretorius and Sonja Bosch

The importance of the semantic annotation of morphological data for agglutinating languages is the departure point of this paper. It discusses the principled extension of the ISOcat data category registry (DCR) to include Zulu morphosyntactic data categories. The focus is on the Zulu noun. Where existing data categories are found appropriate they are used and where new additions are required the published guidelines are followed. The expectation is that these extensions will also be useful for languages that are related to Zulu and share its morphosyntactic structure. The inclusion of the other Zulu word categories forms part of future work.

Understanding Questions and Finding Answers: Semantic Relation Annotation to Compute the Expected Answer Type

Volha Petukhova

This paper presents an annotation scheme for semantic relations used for question classification and answer extraction in an interactive dialogue-based quiz game. The game is concerned with biographical information about famous people's lives, and is often available as unstructured texts on the internet, e.g. the Wikipedia collection (http://www.wikipedia.org). Questions and extracted answers are annotated with dialogue act information using the ISO 24617-2 scheme, and with semantic relations, for which an extensive annotation scheme is developed that combines elements from the TAC KBP slot filling and TREC QA tasks. Dialogue act information, semantic relations and focus words (or word sequences) are used to compute the Expected Answer Type. Th semantic relation annotation scheme is validated according to ISO criteria for the design of a semantic annotation scheme. The tagset is shown to fit the data well.

From Visual Prototypes of Action to Metaphors: Extending the IMAGACT Ontology of Action to Secondary Meanings

Susan Windisch Brown

This paper describes an infrastructure that has been designed to deal with corpus based variations that do not fall within the primary, physical variation of action verbs. We have first established three main categories of secondary variation--metaphor, metonymy and idiom--and criteria for creating types within these categories for each verb. The criteria rely heavily on the images that compose the IMAGACT ontology of action and on widely accepted processes of meaning extension in linguistics. Although figurative language is known for its amorphous, subjective nature, we have endeavored to create a standard, justifiable process for determining figurative language types for individual verbs. We specifically highlight the benefits that IMAGACT’s representation of the primary meanings through videos brings to the understanding and annotation of secondary meanings.

Annotating Cohesion for Multilingual Analysis
Ekaterina Lapshinova-Koltunski and Kerstin Anna Kunz

This paper describes a set of procedures used to semi-automatically annotate a multilingual corpus on the level of cohesion, an important linguistic component of effectively organised and meaningful discourse. The annotation categories we operate with base on different degrees of granularity and account for lexico-grammatical and semantic aspects of different types of cohesion. This annotation scheme allows us to compare and differentiate cohesive features across languages, text types and in written and spoken discourse on different levels of abstraction. Our aim is to obtain a fine-grained and highly precise annotation, at the same time avoiding purely manual annotation. Therefore, we decide for corpus-based semi-automatic procedures to identify candidates expressing cohesion in English and in German. The annotated corpus is one of the few existing resources supporting contrastive studies of cohesion.

Spatio-Temporal Grounding of Claims Made on the Web, in Pheme

Leon Derczynski and Kalina Bontcheva

Social media presents us with a digitally-accessible sample of all human discourse. This sample is full of claims and assertions. While the state of the art in NLP is adapting to the volume, velocity and variety of this sample and the information in it, the accuracy of claims made in social media remain largely unstudied. Pheme, a 36-month EU project started in January 2014, focuses on this fourth challenge: veracity. As a core part of establishing veracity, we need to identify the spatio-temporal context of assertions made on informal websites. This project note introduces the spatio-temporal challenges and planned semantic annotation activities that are part of the Pheme project.

How to Exploit Paralinguistic Features to Identify Acronyms in Texts?

Mathieu Roche

This paper addresses the issue of acronym dictionary building. The first step of the process identifies acronym/definition candidates, the second one selects candidates based on a letter alignment method. This approach has two advantages because it enables (1) to annotate documents, (2) to build specific dictionaries. More precisely, this paper discusses the use of a specific linguistic concept, the gloss, in order to identify candidates. The proposed method based on paralinguistic markers is language-independent.

Lessons Learned from Manual Evaluation of NER Results by Domain Experts

Sungho Shin, Hanmin Jung, Inga Hannemann and Mu Yong Yi

Recently, NER (Named Entity Recognition) has been adopted in many practical areas. People with smart phones would like to have services managing automatically their schedules by scheduling applications having engines inside for extraction of important events to users from texts and emails. Diversifying the application of NER technology to various fields requires higher accuracy this technology. For example, the F-score of our NER system is 0.74 in the laboratory and 0.22 in practice. In order to overcome this issue, NER evaluation should be performed manually, allowing developers or researchers to identify the problems that can occur in practical environment with their

current NER engines in order to improve future versions. This paper addresses the extraction results of NER engines. We approach to the problem by hiring domain experts to evaluate the extraction result. Certain problematic cases that are not expected to be extracted by machines are presented; moreover, feedback from the problems is provided in order to improve the NER engine.

Annotating Discourse Zones in Medical Encounters

Milan Tofiloski, Fred Popowich and Evan Zhang

We improve a visual analytics workflow for analyzing medical interviews by introducing a discourse annotation scheme for creating an effective multi-document visualization that also facilitates inter-document comparison. We introduce the concept of ‘discourse zones’ for bringing together the many disparate terms and concepts used in various research areas. The zones are generalized for usage toward any institutional dyad setting (attorney-witness, teacher-student, physician-patient, etc.), including emergency hotlines and switchboards. Our task involves identifying the medical problems, their solutions, and contexts in medical encounters (e.g. dialogue-based conversations and interviews). The corpora consist of medical interviews between clinicians and caregivers of children with Fetal Alcohol Spectrum Disorder (FASD).

Annotation of Pronouns in a Multilingual Corpus of Mandarin Chinese, English and Japanese

Yu Jie Seah and Francis Bond

A qualitative and quantitative approach was used in this study to examine the distribution of pronouns in three languages,English, Mandarin Chinese and Japanese based on the parallel NTU Multilingual Corpus (NTU-MC). The pronouns are annotated with a componential analysis that allows them to be easily linked across languages. A single text (The Adventure of the Speckled Band, a short story featuring Sherlock Holmes) in three languages is tagged, annotated and linked in the corpus. The results show that although English has the most number of pronouns, Mandarin Chinese has the highest percentage of referential pronouns. Also, English has more translated counterparts in Mandarin Chinese, compared to Japanese. We attributed this to the difference in usage of pronouns in these languages. Deprominalisation, surprisingly, was even for both corpora. Findings from this study can shed some light on translation issues converning pronoun usage for learners of the languages, and may also contribute to improving machine translation of pronouns.

Enriching 'Senso Comune' with Semantic Role Sets

Elisabetta Jezek, Laure Vieu, Fabio Massimo Zanzotto, Guido Vetere, Alessandro Oltramari, Aldo Gangemi and Rossella Varvara

This paper describes the design and the results of a manual annotation methodology for enriching the Senso Comune resource with semantic role sets for predicates. The main issues encountered in applying the 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. We describe the two experiments we carried to verify the reliability of the annotation methodology and to release the annotation scheme. Finally, we discuss the results of the linguistic analysis of the annotated data and report about ongoing work.

Automatic Tagging of Modality: Identifying Triggers and Modal Values

Paulo Quaresma, Amália Mendes, Iris Hendrickx and Teresa Gonçalves

This paper presents an experiment in the automatic tagging of modality in Portuguese. As we are currently lacking a suitable resource with detailed modal information for Portuguese, we experiment with a small sample of 160.000 tokens, manually annotated according to the modality scheme that we previously developed for European Portuguese. We consider modality as the expression of the speaker (or subject's) attitude towards a certain proposition. Our modality scheme accounts for seven major modal values, and nine sub-values. This experiment focuses on three modal verbs which may all have more thatn one modal value: poder (may/can), dever (shall/night) and conseguir (manage to/succeed in/be able to). We rport on the task of correctly detecting the modal uses of poder and dever, since these two verbs may also have non-modalmeanings. For the identification of the modal value of each occurrence of those three verbs, we applied a machine learning approach that takes into consideration all the features available from a syntactic parser's output. We obtained the best performance using SVM with a string kernel, and the system improved the baseline for all three verbs, with a maximum F-score of 76.2.

Using the Crowd to Annotate Metadiscursive Acts

Rui Correia, Nuno Mamede, Jorge Baptista and Maxine Eskenazi

This paper addresses issues relating to the definition and non-expert understanding of metadiscursive acts. We present existing theory on spoken metadiscourse, focusing on one taxonomy that defines metadiscursive concepts in a functional manner, rather than formally. A crowdsourcing annotation task is set up with two main goals: (a) build a corpus of metadiscourse, and (b) assess the understanding of metadiscursive concepts by non-experts. This initial annotation effort focuses on five categories of metadiscourse: INTRODUCING TOPIC, CONCLUDING TOPIC, MARKING ASIDES, EXEMPLIFYING, and EMPHASIZING. The crowdsourcing task is described in detail, including instructions and quality insurance mechanisms. We report results in terms of time-on-task, self-reported confidence, requests for additional context, quantity of occurrences and inter-annotator agreement. Results show the crowd is capable of annotating metadiscourse and give insights into the complexity of the concepts in the taxonomy.

The 10th Workshop on Multimodal Corpora: Combining Applied and Basic Research Targets

May 27th 2014

ABSTRACTS

Editors:

Jens Edlund, Dirk Heylen, Patrizia Paggio

Workshop Programme

09:00 – 09:15 Registration
09:15 – 09:30 Welcome

09:30 – 10:30 Keynote
Hannes Högni Vilhjálmsson: *TBA*

10:30 – 11:00 Coffee break

11:00 – 13:00 Session 1 (Oral)
Masashi Inoue, Toshio Irino, Ryoko Hanada, Nobuhiro Furuyama and Hiroyasu Massaki: *Continuous Annotations for Dialogue Status and Their Change Points*
Eli Pincus and David Traum: *Towards a Multimodal Taxonomy of Dialogue Moves for Word-Guessing Games*
Bayu Rahayudi, Ronald Poppe and Dirk Heylen: *Gaze Patterns in the Twente Debate Corpus*
Kirsten Bergmann, Ronald Böck and Petra Jaecks: *EmoGest: Investigating the Impact of Emotions on Spontaneous Co-speech Gestures*

13:00 – 14:00 Lunch break

14:00 – 15:00 Session 2 (Oral)
Ivan Gris, David Novick, Mario Gutierrez and Diego Rivera: *The "Vampire King" (Version 2) Corpus*
Kristina Nilsson Björkenstam and Mats Wirén: *Multimodal Annotation of Synchrony in Longitudinal Parent-Child Interaction*

15:00 – 16:00 Session 3 (Poster)
Federica Cavicchio, Amanda Brown, Reyhan Furman, Shanley Allen, Asli Özyürek, Tomoko Ishizuka and Sotaro Kita: *Annotation of space and manner/path configuration in bilinguals' speech and manual gestures*
Oliver Schreer and Stefano Masneri: *Automatic Video Analysis for Annotation of Human Body Motion in Humanities Research*
Trine Eilersen and Costanza Navarretta: *A Multimodal Corpus of Communicative Behaviors of Disabled Individuals during HRI*
Jens Edlund, Mattias Heldner and Marcin Wlodarczak: *Catching wind of multiparty conversation*

16:00 – 16:30 Coffee break

16:30 – 10:30 Session 4 (Oral)
Gabriel Murray: *Resources for Analyzing Productivity in Group Interactions*
Nesrine Fourati, Jing Huang and Catherine Pelachaud: *Dynamic stimuli visualization for experimental studies of body language*

17:30 – 18:00 Business meeting, close

Workshop Organizers

Jens Edlund	KTH Royal Institute of Technology, Sweden
Dirk Heylen	University of Twente, The Netherlands
Patrizia Paggio	University of Copenhagen, Denmark/University of Malta, Malta

Workshop Programme Committee

Jens Allwood	University of Gothenburg, Sweden
Susanne Burger	Carnegie Mellon University, USA
Jens Edlund	KTH Royal Institute of Technology, Sweden
Dirk Heylen	University of Twente, The Netherlands
Costanza Navarretta	University of Copenhagen, Denmark/University of Malta, Malta
David Novick	University of Texas at El Paso, USA
Patrizia Paggio	University of Copenhagen, Denmark
Ronald Poppe	University of Twente, The Netherlands
Albert Ali Salah	Boğaziçi University, Turkey
David Schlangen	Bielefeld University, Germany
David Traum	Institute for Creative Technologies, USA

Introduction

We are pleased that in 2014, the 10th Workshop on Multimodal Corpora is once again returning home and is collocated with LREC, this time in Reykjavik, Iceland. The workshop follows in a series previously held at LREC 2000, 2002, 2004, 2006, 2008, 2010; at ICMI 2011; at LREC 2012; and at IVA 2013 (all workshops of the series are documented under www.multimodal-corpora.org).

As always, we present a wide cross-section of the field, with contributions ranging from collection efforts, coding, validation and analysis methods, to tools and applications of multimodal corpora. Given that LREC this year emphasizes the use of corpora to solve language technology problems and develop useful applications and services, we aim for this workshop also to highlight the usefulness of multimodal corpora to applied research as well as basic research. Many of the unimodal speech corpora collected over the past decades have served a double purpose: on the one hand, they have enlightened our view on the basic research question of how speech works and how it is used; on the other hand, they have forwarded the applied research goal of developing better speech technology applications. This reflects the dual nature of speech technology, where funding demands often require researchers to follow research agendas that target applied and basic research goals in parallel.

Multimodal corpora are potentially more complex than unimodal corpora, and their design poses an even greater challenge. Yet the benefits to be gained from designing with a view to both applied and basic research remain equally desirable. Against this background, the theme for this instalment of Multimodal Corpora is how multimodal corpora can be designed to serve this double purpose.

Session 1

Tuesday May 27th 11:00 – 13:00
Chairperson: TBA

Continuous Annotations for Dialogue Status and Their Change Points

Masashi Inoue, Toshio Irino, Ryoko Hanada, Nobuhiro Furuyama and Hiroyasu Massaki

This paper presents an attempt to continuously annotate the emotion and status of multimodal corpora for understanding psychotherapeutic interviews. The collected continuous annotations are then used as the signal data to find change points in the dialogues. Our target dialogues are carried between clients with some psychological problems and their therapists. We measured two values, namely the degree of the dialogue progress and the degree of clients being listened to. The first value reflects the goal-oriented nature of the target dialogues. The second value corresponds to the idea of active listening that is considered as an important aspect in psychotherapy. We have modified an existing continuous emotion annotation toolkit that has been created for tracking generic emotion of dialogues. By applying a change point detection algorithm on the obtained annotations, we evaluated the validity and utility of the collected annotation based on our method.

Towards a Multimodal Taxonomy of Dialogue Moves for Word-Guessing Games

Eli Pincus and David Traum

We develop a taxonomy for guesser and clue-giver dialogue moves in word guessing games. The taxonomy is designed to aid in the construction of a computational agent capable of participating in these games. We annotate the word guessing game of the multimodal Rapid Dialogue Game (RDG) corpus, RDG-Phrase, with this scheme. The scheme classifies clues, guesses, and other verbal actions as well as non-verbal actions such as gestures into different types. Cohen kappa inter-annotator agreement statistics for clue/non-clue and guess/non-guess are both approximately 76%, and the kappas for clue type and guess type are 59% and 75%, respectively. We discuss phenomena and challenges we encounter during annotation of the videos such as co-speech gestures, gesture disambiguation, and gesture discretization.

Gaze Patterns in the Twente Debate Corpus

Bayu Rahayudi, Ronald Poppe and Dirk Heylen

Different patterns of verbal and nonverbal behaviours have been associated with turn-taking in face-to-face conversations. Gaze is one that has been studied extensively. An important factor that determines the exact patterns in a particular conversation is the nature of the conversation; whether it is dyadic or multi-party, whether it is a chat or a heated debate, etcetera. In this paper we present a first analysis of the gaze patterns in the Twente Debate Corpus to investigate how the particular setting that was chosen influences the patterns in gaze behaviour. This analysis is meant to provide us with better insight in the features that are needed to improve automatic prediction algorithms such as those that predict the end of a turn.

EmoGest: Investigating the Impact of Emotions on Spontaneous Co-speech Gestures

Kirsten Bergmann, Ronald Böck and Petra Jaecks

Spontaneous co-speech gestures are an integral part of human communicative behavior. Little is known, however, about how they reflect a speaker's emotional state. In this paper, we describe the setup of a novel body movement database. 32 participants were primed with emotions (happy, sad, neutral) by listening to selected music pieces and, subsequently, fulfilled a gesture-eliciting task. We present our methodology of evaluating the effects of emotion priming with standardized questionnaires, and via automatic emotion recognition of the speech signal. First results suggest that emotional priming was successful, thus, paving the way for further analyses comparing the gestural behavior across the three experimental conditions.

Session 2

Tuesday May 27th 14:00 – 15:00
Chairperson: TBA

The "Vampire King" (Version 2) Corpus

Ivan Gris, David Novick, Mario Gutierrez and Diego Rivera

As part of a study examining nonverbal and paralinguistic behaviors in conversations between humans and embodied conversational agents (ECAs), we collected a corpus of human subjects interacting with an ECA in an adventure game. In the interaction, the ECA served as a narrator for a game entitled "Escape from the Castle of the Vampire King," which was inspired by text-based computer games such as Zork. The corpus described here is based on Version 2 of the game, in which a map of the castle was displayed on the wall behind the ECA. The system was not a Wizard-of-Oz simulation; the system responded using speech recognition and utterance generation. The experiment had two conditions, familiar and non-familiar, defined by the degree of nonverbal extraversion presented by the ECA. The corpus includes 20 subjects, each of whom interacted with the game for 30-minute sessions on two consecutive days, for a total of approximately 1200 minutes of interaction. All 40 sessions were both audiovisually recorded and automatically annotated for speech and basic posture using a Kinect sensor. The corpus includes (a) the automated annotations for speech and posture and (b) manual annotations for gaze, nods and interrupts.

Multimodal Annotation of Synchrony in Longitudinal Parent-Child Interaction

Kristina Nilsson Björkenstam and Mats Wirén

This paper describes the multimodal annotation of speech, gaze and hand movement in a corpus of longitudinal parent–child interaction, and reports results on synchrony, structural regularities which appear to be a key means for parents to facilitate learning of new concepts to children. The results provide additional support for our previous finding that parents display decreasing synchrony as a function of the age of the child.

Session 3

Tuesday May 27th 15:00 – 16:00
Chairperson: TBA

Annotation of space and manner/path configuration in bilinguals' speech and manual gestures

Federica Cavicchio, Amanda Brown, Reyhan Furman, Shanley Allen, Asli Özyürek, Tomoko Ishizuka and Sotaro Kita

Different languages and cultures use gestures differently. The goal of this paper is describing the coding scheme used to annotate a corpus of English/Italian bilinguals and English and Italian monolinguals describing, a set of stimuli designed to elicit the description of manner and path events and the corresponding gestures ("the Tomato Man stimuli"). The first question we investigated was the relationship between clause structure for motion event expressions and gestural representation of the same event. From the seminal work of Kita and Özyürek (2003), many studies have investigated manner and path in the verbalization of motion events and the co-produced manual gestures in different languages. Following Talmy's (1985) typology, English allows verbal constructions to conflate complex meaning within a single clause as path can be expressed as a "satellite" to the verb. That is, manner and path can be expressed in within a single clause (e.g. roll down). On the other hand, Italian is more restricted in situations in which manner of motion verbs can occur with path phrases. That is, manner and path are often realized by two verbs (scende rotolando, i.e. it goes down rolling). We describe the annotation scheme we used to code speech and gesture in English/Italian bilinguals and monolinguals. Three annotators codified 403 tokens corresponding to the gesture stroke phase. We coded gestural and verbal expressions of manner and path and gesture space.

Automatic Video Analysis for Annotation of Human Body Motion in Humanities Research

Oliver Schreer and Stefano Masneri

The analysis of multi-modal audio-visual data is the very first step to perform research on gestural behaviour in a variety of disciplines such as psycho-linguistic, psychology, behaviour analysis or sign language. The annotation of human gestures and motion of hands is a very time consuming process and requests a lot of effort in terms of personal. Due to the large amount of available data in existing corpora, much video material has not been annotated and even not touched, i.e. valuable material cannot be exploited for research. Thanks to modern video processing algorithms some of the most time consuming annotation tasks can be performed automatically. In this paper, we present a video analysis tool that is specifically designed to perform automatic analysis and annotation of video material for the above mentioned research domains. The presented algorithm provides a large variety of annotations required for behaviour analysis without any user interaction in a fully automatic way. The proposed video analysis tool is currently designed to provide annotations according to the NEUROGES coding system for gestural behaviour, but it can provide also other means of annotations for other coding schemes.

A Multimodal Corpus of Communicative Behaviors of Disabled Individuals during HRI

Trine Eilersen and Costanza Navarretta

This paper describes the collection of a multimodal corpus of video- and audio-recorded interactions between an anthropomorphic robot and normal and cognitively disabled individuals. The aim of the work is to provide data for the study of the multimodal behaviors of the two groups of test participants during the conversational interactions. The study of the communicative multimodal behaviors of possible future users of assistive social robotics is expected to provide useful information for the development of human-robot interfaces. The data-collection was conducted using an anthropomorphic robot, NAO, which was interacting with the participants via a Wizard of Oz setting and a modular dialogue-script. Test participants were recruited from 2 municipal care centers for adult individuals with multiple disabilities located in Copenhagen, Denmark. The corpus was collected over 2 weeks and consists of recordings of 17 dyadic interactions. Each interaction comprises chat-based conversation units and a small cooperative game with a ball. A first analysis of the video-recordings shows that both disabled and non-disabled participants interacted multimodally (speech and body behaviors) with the robot. Furthermore, the answers of the participants to a questionnaire about their feelings towards the communicative situation show that they were unaffected by the experimental set-up while they were very affected by the meeting with the robot.

Catching wind of multiparty conversation

Jens Edlund, Mattias Heldner and Marcin Wlodarczak

The paper describes the design of a novel corpus of respiratory activity in spontaneous multiparty face-to-face conversations in Swedish. The corpus is collected with the primary goal of investigating the role of breathing for interactive control of interaction. Physiological correlates of breathing are captured by means of respiratory belts, which measure changes in cross sectional area of the rib cage and the abdomen. Additionally, auditory and visual cues of breathing are recorded in parallel to the actual conversations. The corpus allows studying respiratory mechanisms underlying organisation of spontaneous communication, especially in connection with turn management. As such, it is a valuable resource both for fundamental research and speech techonology applications.

Session 4

Tuesday May 27th 16:30 – 17:30
Chairperson: TBA

Resources for Analyzing Productivity in Group Interactions

Gabriel Murray

Productivity can vary both within and across meetings. In this work, we consider the question of how to measure productivity, and survey some of the available and potential resources that correspond to productivity. We then describe an initial experiment in which we define productivity in terms of the percentage of sentences from a meeting that are considered summary-worthy. Given that simple definition of productivity, we fit a logistic regression model to predict productivity levels of meetings using linguistic and structural features.

Dynamic stimuli visualization for experimental studies of body language

Nesrine Fourati, Jing Huang and Catherine Pelachaud

Understanding human body behavior have relied on perceptive studies. Lately, several experimental studies have been conducted with virtual avatars that reproduce human body movements. The visualization of human body behaviors stimuli, using avatars, may introduce bias for human perception comprehension. Indeed, the choice of the virtual camera trajectory and orientation affects the display of the stimuli. In this paper, we propose control functions for the virtual camera.

3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing

27 May 2014

ABSTRACTS

Editors:

Christian Chiarcos, John Philip McCrae,Petya Osenova, Cristina Vertan

Workshop Programme*]*

08:30 - 09:00 – Opening and Introduction by Workshop Chair(s)

09:00 – 10:00 – Invited Talk
Piek Vossen, *The Collaborative Inter-Lingual-Index for harmonizing wordnets*

10:00 – 10:30 – Session 1: Modeling Lexical-Semantic Resources with *lemon*
Andon Tchechmedjiev, Gilles Sérasset, Jérôme Goulian and Didier Schwab, *Attaching Translations to Proper Lexical Senses in DBnary*

10:30 – 11:00 Coffee break

11:00-11:20– Session 1: Modeling Lexical-Semantic Resources with *lemon*
John Philip McCrae, Christiane Fellbaum and Philipp Cimiano, *Publishing and Linking WordNet using lemon and RDF*

11:20-11:40– Session 1: Modeling Lexical-Semantic Resources with *lemon*
Andrey Kutuzov and Maxim Ionov*, Releasing genre keywords of Russian movie descriptions as Linguistic Linked Open Data: an experience report*

11:40-12:00– Session 2: Metadata
Matej Durco and Menzo Windhouwer, *From CLARIN Component Metadata to Linked Open Data*

12:00-12:20– Session 2: Metadata
Gary Lefman, David Lewis and Felix Sasaki*, A Brief Survey of Multimedia Annotation Localisation on the Web of Linked Data*

12:20-12:50– Session 2: Metadata
Daniel Jettka, Karim Kuropka, Cristina Vertan and Heike Zinsmeister, *Towards a Linked Open Data Representation of a Grammar Terms Index*

12:50-13:00 – Poster slam – Data Challenge

13:00 – 14:00 Lunch break

14:00 – 15:00 – Invited Talk
Gerard de Mello, *From Linked Data to Tightly Integrated Data*

15:00 – 15:30 – Section 3: Crosslinguistic Studies
Christian Chiarcos and Maria Sukhareva, Linking Etymological Databases. *A case study in Germanic*

15:30 – 16:00 – Section 3: Crosslinguistic Studies
Fahad Khan, Federico Boschetti and Francesca Frontini, *Using lemon to Model Lexical Semantic Shift in Diachronic Lexical Resources*

16:00 – 16:30 Coffee break

16:30 – 17:00 – Section 3: Crosslinguistic Studies
Steven Moran and Michael Cysouw, *Typology with graphs and matrices*

17:00 – 17:30 – Section 3: Crosslinguistic Studies
Robert Forkel, *The Cross-Linguistic Linked Data project*

17:30 – 18:30 – Poster Session – Data Challenge
Gilles Sérasset and Andon Tchechmedjiev*, Dbnary: Wiktionary as Linked Data for 12 Language Editions with Enhanced Translation Relations*

Maud Ehrmann, Francesco Cecconi, Daniele Vannella, John Philip McCrae, Philipp Cimiano and Roberto Navigli, *A Multilingual Semantic Network as Linked Data: lemon-BabelNet*

Gabriela Vulcu, Raul Lario Monje, Mario Munoz, Paul Buitelaar and Carlos A. Iglesias, *Linked-Data based Domain-Specific Sentiment Lexicons*

Tomáš Kliegr, Vaclav Zeman and Milan Dojchinovski, *Linked Hypernyms Dataset - Generation framework and Use Cases*

Ismail El Maarouf, Jane Bradbury and Patrick Hanks*, PDEV-lemon: a Linked Data implementation of the Pattern Dictionary of English Verbs based on the Lemon model*

18:30 – 19:00 – Discussions and Closing

Workshop Organizers

Christian Chiarcos	Goethe-University Frankfurt am Main, Germany
John Philip McCrae	University of Bielefeld, Germany
Kiril Simov	Bulgarian Academy of Sciences, Sofia, Bulgaria
Antonio Branco	University of Lisbon, Portugal
Nicoletta Calzolari	ILC-CNR, Italy
Petya Osenova	University of Sofia, Bulgaria
Milena Slavcheva	JRC-Brussels, Belgium
Cristina Vertan	University of Hamburg, Germany

Workshop Programme Committee

Eneko Agirre	University of the Basque Country, Spain
Guadalupe Aguado	Universidad Politécnica de Madrid, Spain
Peter Bouda	Interdisciplinary Centre for Social and Language Documentation, Portugal
Steve Cassidy	Macquarie University, Australia
Damir Cavar	Eastern Michigan University, USA
Walter Daelemans	University of Antwerp, Belgium
Ernesto William De Luca	University of Applied Sciences Potsdam, Germany
Gerard de Melo	University of California at Berkeley, USA
Dongpo Deng	Institute of Information Sciences, Academia Sinica, Taiwan
Alexis Dimitriadis	Universiteit Utrecht, The Netherlands
Jeff Good	University at Buffalo, USA
Asunción Gómez Pérez	Universidad Politécnica de Madrid, Spain
Jorge Gracia	Universidad Politécnica de Madrid, Spain
Walther v. Hahn	University of Hamburg, Germany
Eva Hajicova	Charles University Prague, Czech Republic
Harald Hammarström	Radboud Universiteit Nijmegen, The Netherlands
Yoshihiko Hayashi	Osaka University, Japan
Sebastian Hellmann	Universität Leipzig, Germany
Dominic Jones	Trinity College Dublin, Ireland
Lutz Maicher	Universität Leipzig, Germany
Pablo Mendes	Open Knowledge Foundation Deutschland, Germany
Steven Moran	Universität Zürich, Switzerland/Ludwig Maximilian University, Germany
Sebastian Nordhoff	Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
Maciej Piasecki	Wroclaw University of Technology, Poland
Adam Przepiorkowski	IPAN, Polish Academy of Sciences, Poland
Laurent Romary	INRIA, France
Felix Sasaki	DFKI, Germany Intelligenz, Germany

Session Introduction

Tuesday 27 May,08:30 – 09:00
Chairperson: Nicoletta Calzolari

Linked Data in Linguistics 2014: Introduction and Overview
Christian Chiarcos, John McCrae, Petya Osenova and Cristina Vertan

Abstract
The Linked Data in Linguistics (LDL) workshop series brings together researchers from various fields of linguistics, natural language processing, and information technology to present and discuss principles, case studies, and best practices for representing, publishing and linking linguistic data collections. A major outcome of our work is the Linguistic Linked Open Data (LLOD) cloud, an LOD (sub-)cloud of linguistic resources, which covers various linguistic data bases, lexicons, corpora, terminology and metadata repositories.
As a general introduction into the topic, we describe the concept of Linked Data, its application in linguistics and the development of the Linguistic Linked Open Data (LLOD) cloud since LDL-2013. We present the contributions of LDL-2014, the associated data challenge and its results and present the newly compiled LLOD cloud diagram.
The third instantiation of this series, collocated with the 9th Language Resources and Evaluation Conference (LREC-2014), May 27th, 2014, in Reykjavik, Iceland, is specifically dedicated to the study of Multilingual Knowledge Resources and Natural Language Processing, although contributions with respect to any application of Linked Data to linguistically and/or NLP-relevant resources are welcome, as well.

Session Invited Talks

Tuesday 27 May,09:00 – 10:00 and 14:00 – 15:00
Chairperson: Petya Osenova

The Collaborative Inter-Lingual-Index for harmonizing wordnets
Piek Vossen

Abstract
The EuroWordNet project proposed an Inter-Lingual-Index (ILI) to link independently developed wordnets. The ILI was initially filled with the English WordNet. Since then many wordnets have been developed following this model but no work has been done on the ILI since.

At the last Global Wordnet Conference in Tartu (2014), we decided to take up the initial ideas from EuroWordNet and establish a ILI platform that will result in a fund of concepts and meanings that is not just dependent on English.

This concept repository will be published as Linked Open Data with a collaborative social platform to add new concepts and link synsets from different wordnets. In this way, we can match synsets across wordnets even if there is no English equivalent. Modifications and changes are reported back to the community and feedback is given on 'semantic impact' of changes.

The ILI supports a harmonization process for wordnets. It will allow us to flesh out differences in lexicalizations across languages. As proposed in EuroWordNet, the conceptual index can also be formalized by linking ontologies to these concepts, such as SUMO, DOLCE or DBPedia. The project seeks to establish a semantic layer for interpretation of text across languages. In a number of projects, we develop deep-reading technologies to extract information from texts across different languages. Such projects directly benefit from ILI.

As an example, we explain how we were able to do semantic-role-labelling in Dutch using the SemLink mappings for English that were transferred to the Dutch wordnet.

From Linked Data to Tightly Integrated Data
Gerard de Mello

Abstract
The ideas behind the Web of Linked Data have great allure. Apart from the prospect of large amounts of freely available data, we are also promised nearly effortless interoperability. Common data formats and protocols have indeed made it easier than ever to obtain and work with information from different sources simultaneously, opening up new opportunities in linguistics, library science, and many other areas.
In this talk, however, I argue that the true potential of Linked Data can only be appreciated when extensive cross-linkage and integration engenders an even higher degree of interconnectedness. This can take the form of shared identifiers, e.g. those based on Wikipedia and WordNet, which can be used to describe numerous forms of linguistic and commonsense knowledge. An alternative is to rely on sameAs and similarity links, which can automatically be discovered using scalable approaches like the LINDA algorithm but need to be interpreted with great care, as we have observed in experimental studies. A closer level of linkage is achieved when resources are also connected at the taxonomic level, as exemplified by the MENTA approach to taxonomic data integration. Such integration means that one can buy into ecosystems already carrying a range of valuable pre-existing assets. Even more tightly integrated resources like Lexvo.org combine triples from multiple sources into unified, coherent knowledge bases.
Finally, I also comment on how to address some remaining challenges that are still impeding a more widespread adoption of Linked Data on the Web. In the long run, I believe that such steps will lead us to significantly more tightly integrated Linked Data.

Session Modelling Lexical-Semantic Resources with *lemon*

Tuesday 27 May, 10:00 – 11:40
Chairperson: Christian Chiarcos

Attaching Translations to Proper Lexical Senses in DBnary
Andon Tchechmedjiev, Gilles Sérasset, Jérôme Goulian and Didier Schwab

Abstract
The DBnary project aims at providing high quality Lexical Linked Data extracted from different Wiktionary language editions. Data from 10 different languages is currently extracted for a total of over 3.16M translation links that connect lexical entries from the 10 extracted languages, to entries in more than one thousand languages. In Wiktionary, glosses are often associated with translations to help users understand to what sense they refer to, wither though a textual definition or a target sense number. In this article we aim at the extraction of as much of this information as possible and then the disambiguation of the corresponding translations for all languages available. We use an adaptation of various textual and semantic similarity techniques based on partial or fuzzy gloss overlaps to disambiguate the translation relations (To account for the lack of normalization, e.g. lemmatization and PoS tagging) and then extract some of the sense number information present to build a gold standard so as to evaluate our disambiguation as well as tune and optimize the parameters of the similarity measures. We obtain F-measures of the order of 80% (on par with similar work on English only), across the three languages where we could generate a gold standard (French, Portuguese, Finnish) and show that most of the disambiguation errors are due to inconsistencies in Wiktionary itself that cannot be detected at the generation of DBNary (shifted sense numbers, inconsistent glosses, etc.).

Publishing and Linking WordNet using lemon and RDF
John Philip McCrae, Christiane Fellbaum and Philipp Cimiano

Abstract
In this paper we provide a description of a dataset consisting of data from the Princeton WordNet. This version is intended to provide canonical URIs that can be used by a wide variety of lexical resources to express their linking as part of the Linguistic Linked Open Data Cloud. Furthermore, this is the first version to use the lemon model and we describe how we represent WordNet with this model.

Releasing genre keywords of Russian movie descriptions as Linguistic Linked Open Data: an experience report
Andrey Kutuzov and Maxim Ionov

Abstract
This paper describes a lexical database derived from a larger dataset of movie semantic properties. The larger dataset is a collection of RDF triples from most popular Russian movie delivery sites (mostly Schema.org predicates). The movies in these triples were classified according to their respective genre, and then keywords which are characteristic to descriptions of such movies, were extracted using log-likelihood approach. The set of such keywords (about 8000 lemmas) with various keyness attributes is published as Linguistic Linked Open Data, with the help of Lemon model. Additionally, lemmas were linked to Russian DBPedia Wiktionary.

Session Metadata
Tuesday 27 May, 11:40 – 12:50
Chairperson: Paul Cimiano

CLARIN Component Metadata to Linked Open Data
Matej Durco and Menzo Windhouwer

Abstract
In the European CLARIN infrastructure a growing number of resources are described with Component Metadata. In this paper we describe a transformation to make this metadata available as linked data. After this first step it becomes possible to connect the CLARIN Component Metadata with other valuable knowledge sources in the Linked Data Cloud.

A Brief Survey of Multimedia Annotation Localisation on the Web of Linked Data
Gary Lefman, David Lewis and Felix Sasaki

Abstract
Multimedia annotation generates a vast amount of monolingual data that can help to describe audio, video, and still images. These annotations are, however, unlikely to be useful to people that cannot communicate through the same language. Annotations may also contain insufficient context for people coming from different cultures, so there is a need for localised annotations, in addition to localised multimedia. We have performed a brief survey of multimedia annotation capabilities, choosing Flickr as a representative candidate of open social media platforms. The focus of our examination was on the role of localisation in multimedia ontologies and Linked Data frameworks. In order to share annotated multimedia effectively on the Web of Linked Data, we believe that annotations should be linked to similar resources that have already been adapted for other languages and cultures. In the absence of a Linked Data framework, monolingual annotations remain trapped in silos and cannot, therefore, be shared with other open social media platforms. This discovery led

to the identification of a gap in the localisation continuity between multimedia annotations and the Web of Linked Data.

Towards a Linked Open Data Representation of a Grammar Terms Index
Daniel Jettka, Karim Kuropka, Cristina Vertan and Heike Zinsmeister

Abstract
In this paper we report on-going work on the creation of HyperGramm, a Linked Open Data set of German grammar terms. It is based on a print-oriented manually created resource, which contains different types of internal and external linking relations that are either explicitly marked by formatting or only implicitly encoded in the language. The initial aim of the HyperGramm resource was the on-line visualization of the terms. Since this resource can be used in a series of other scenarios both for research and learning purposes, it is desirable that the representation captures as much information as possible about the internal structure of the original resource. We first motivate a conversion into an intermediate well-defined XML presentation, which itself serves as basis for the RDF modeling. Then we detail the RDF model and demonstrate how it allows us to encode the internal structure and the linking mechanisms in an explicit and interoperable way. In addition, we discuss a possible integration of HyperGramm in the LOD Cloud.

Session Crossslinguistic Studies

Tuesday 27 May, 15:00 – 17:30
Chairperson: Cristina Vertan

Linking Etymological Databases. A case study in Germanic
Christian Chiarcos and Maria Sukhareva

Abstract
This paper deals with resources for the study of older Germanic languages currently developed at the Goethe- University Frankfurt am Main, Germany. Here, we focus on etymological dictionaries that provide explicit information about diachronic phonological correspondences between lexemes at different language stages, and describe pilot studies on (a) their modeling in RDF, (b) their linking with other resources, and (c) possible applications of the resulting resource. To our best knowledge, this data represents the first attempt to bring together etymological databases with the world of (Linguistic) Linked Open Data. This is surprising, as the application of the Linked Data paradigm in this domain is particularly promising, as the basic nature of etymology involves cross-references between different languagespecific dictionaries.

Using lemon to Model Lexical Semantic Shift in Diachronic Lexical Resources
Fahad Khan, Federico Boschetti and Francesca Frontini

Abstract
In this paper we propose a model, called lemonDIA for representing lexical semantic change using the lemon framework and based on the notion of perdurant. Namely we extend the notion of sense in lemon by adding a temporal dimension and then define a class of perdurant entities that represents a shift in meaning of a word and which contains different related senses. We start by discussing the general problem of semantic shift and the utility of being able to easily access and represent such information in diachronic lexical resources. We then describe our model and illustrate it with examples.

Typology with graphs and matrices
Steven Moran and Michael Cysouw

Abstract
In this paper we show how the same data source can be represented in three different data formats -- graphs, tables and matrices. After extracting table data from aggregated graphs data sources in the Linguistic Linked Open Data cloud, we convert these tables into numerical matrices to which we can apply mathematical formulations of linear algebra. As one example of the application of matrix algebra for language comparison, we identify clusters of association between disparate typological databases by leveraging the transformation of different data formats and Linked Data.

The Cross-Linguistic Linked Data project
Robert Forkel

Abstract
The Cross-Linguistic Linked Data project (CLLD) helps record the world's language diversity heritage by establishing an interoperable data publishing infrastructure. I describe the project and the environment it operates in, with an emphasis on the datasets that are published within the project. The publishing infrastructure is built upon a custom software stack -- the clld framework -- which is described next. I then proceed to explain how Linked Data plays an important role in the strategy regarding interoperability and sustainability. Finally I gauge the impact the project may have on its environment.

Session Poster session – Data Challenge

Tuesday 27 May, 17:30 – 18:30
Chairperson: John McCrae

Dbnary: Wiktionary as Linked Data for 12 Language Editions with Enhanced Translation Relations
Gilles Sérasset and Andon Tchechmedjiev

Abstract
After winning the Monnet Challenge in 2012, we continued our efforts in extracting multilingual wiktionary data. This data, made available as Linked Data structured using the LEMON Model, now contains 12 language editions. This short paper presents the current status of the dbnary dataset.
The extracted data is registered at \url{http://thedatahub.org/dataset/dbnary}. Explanations, statistics and data may be accessed via the dataset web site: \url{http://kaiko.getalp.org/about-dbnary/}.

A Multilingual Semantic Network as Linked Data: lemon-BabelNet
Maud Ehrmann, Francesco Cecconi, Daniele Vannella, John Philip McCrae, Philipp Cimiano and Roberto Navigli

Abstract
Empowered by Semantic Web technologies and the recent Linked Data uptake, the publication of linguistic data collections on the Web is, apace with the Web of Data, encouragingly progressing. Indeed, with its long-standing tradition of linguistic resource creation and handling, the Natural Language Processing community can, in many respects, benefit greatly from the Linked Data paradigm. As part of our participation to the Data Challenge associated to the Linked Data in

Linguistics Workshop, this paper describes the lemon-BabelNet dataset, a multilingual semantic network published as Linked Data.

Linked-Data based Domain-Specific Sentiment Lexicons
Gabriela Vulcu, Raul Lario Monje, Mario Munoz, Paul Buitelaar and Carlos A. Iglesias,

Abstract
In this paper we present a dataset composed of domain-specific sentiment lexicons in six languages for two domains. We used existing collections of reviews from TripAdvisor, Amazon, the Stanford Network Analysis Project and the OpinRank Review Dataset. We use an RDF model based on the lemon and Marl formats to represent the lexicons. We describe the methodology that we applied to generate the domain-specific lexicons and we provide access information to our datasets.

Linked Hypernyms Dataset - Generation framework and Use Cases
Tomáš Kliegr, Vaclav Zeman and Milan

Abstract
The Linked Hypernyms Dataset (LHD) provides entities described by Dutch, English and German Wikipedia articles with types taken from the DBpedia namespace. LHD contains 2.8 million entity-type assignments. Accuracy evaluation is provided for all languages. These types are generated based on one-word hypernym extracted from the free text of Wikipedia articles, the dataset is thus to a large extent complementary to DBpedia 3.8 and YAGO 2s ontologies. LHD is available at http://ner.vse.cz/datasets/linkedhypernyms.

PDEV-lemon: a Linked Data implementation of the Pattern Dictionary of English Verbs based on the Lemon model
Ismail El Maarouf, Jane Bradbury and Patrick Hanks

Abstract
PDEV-Lemon is the Linguistic Linked Data resource built from PDEV (Pattern Dictionary of English Verbs), using the Lemon lexicon model (McCrae et al., 2011). PDEV is a dictionary which provides insight into how verbs collocate with nouns and words using an empirically well-founded apparatus of syntactic and semantic categories. It is a valuable resource for Natural Language Processing because it specifies in detail the contextual conditions that determine the meaning of a word. Over 1000 verbs have been analysed to date. PDEV-Lemon is built using the Lemon model, the LExicon Model for ONtologies.

Building and Using Comparable Corpora

Date: 27 May 2014

ABSTRACTS

Editors:
Pierre Zweigenbaum, Serge Sharoff, Reinhard Rapp,
Ahmet Aker, Stephan Vogel

Workshop Programme

Session Opening: (9:00-10:00) Invited talk

09:00–10:00 *Crowdsourcing Translation*
Chris Callison-Burch

Session B: (10:00-12:30) Building corpora

10:00–10:30 *Construction of a French-LSF corpus*
Michael Filhol and Xavier Tannier

10:30–11:00 *Merging Comparable Data Sources for the Discrimination of Similar Languages: The DSL Corpus Collection*
Marcos Zampieri, Nikola Ljubesic and Jorg Tiedemann

10:30–11:00 **Coffee break**

11:30–12:00 *Building comparable corpora from social networks*
Marwa Trabelsi, Malek Hajjem and Chiraz Latiri

12:00–12:30 *Twitter as a Comparable Corpus to build Multilingual Affective Lexicons*
Amel Fraisse and Patrick Paroubek

12:30–14:00 **Lunch break**

Session MT: (14:00-16:00) Machine Translation

14:00–14:30 *Comparabilty of Corpora in Human and Machine Translation*
Haithem Afli, Loïc Barrault and Holger Schwenk

14:30–15:00 *Extended Translation Memories for Multilingual Document Authoring*
Jean-Luc Meunier and Marc Dymetman

15:00–15:30 *Using partly multilingual patents to support research on multilingual IR by building translation memories and MT systems*
Lingxiao Wang, Christian Boitet and Mathieu Mangeot

15:30–16:00 *Comparability of Corpora in Human and Machine Translation*
Ekaterina Lapshinova-Koltunski and Santanu Pal

16:00–16:30 **Coffee break**

Session T: (16:30-17:30) Terminology

16:30–17:00 *Identifying Japanese-Chinese Bilingual Synonymous Technical Terms from Patent Families*
Zi Long, Lijuan Dong, Takehito Utsuro, Tomoharu Mitsuhashi and Mikio Yamamoto

17:00–17:30 *Revisiting comparable corpora in connected space*
Pierre Zweigenbaum

Session P: (17:30-18:00) Panel on a shared task

Workshop Organising Committee:

Pierre Zweigenbaum, LIMSI, CNRS, Orsay, France (Chair)
Serge Sharoff, University of Leeds, UK
Reinhard Rapp, Universities of Mainz, Germany, and Aix-Marseille, France
Ahmet Aker, University of Sheffield, UK
Stephan Vogel, QCRI, Qatar

Workshop Programme Committee:

Ahmet Aker (University of Sheffield, UK)
Srinivas Bangalore (AT&T Labs, US)
Caroline Barrière (CRIM, Montréal, Canada)
Chris Biemann (TU Darmstadt, Germany)
Hervé Déjean (Xerox Research Centre Europe, Grenoble, France)
Kurt Eberle (Lingenio, Heidelberg, Germany)
Andreas Eisele (European Commission, Luxembourg)
Éric Gaussier (Université Joseph Fourier, Grenoble, France)
Gregory Grefenstette (Exalead, Paris, France)
Silvia Hansen-Schirra (University of Mainz, Germany)
Hitoshi Isahara (Toyohashi University of Technology)
Kyo Kageura (University of Tokyo, Japan)
Adam Kilgarriff (Lexical Computing Ltd, UK)
Natalie Kübler (Université Paris Diderot, France)
Philippe Langlais (Université de Montréal, Canada)
Emmanuel Morin (Université de Nantes, France)
Dragos Stefan Munteanu (Language Weaver, Inc., US)
Lene Offersgaard (University of Copenhagen, Denmark)
Ted Pedersen (University of Minnesota, Duluth, US)
Reinhard Rapp (Université Aix-Marseille, France)
Sujith Ravi (Google, US)
Serge Sharoff (University of Leeds, UK)
Michel Simard (National Research Council Canada)
Richard Sproat (OGI School of Science & Technology, US)
Tim Van de Cruys (IRIT-CNRS, Toulouse, France)
Stephan Vogel, QCRI (Qatar)
Guillaume Wisniewski (Université Paris Sud & LIMSI-CNRS, Orsay, France)
Pierre Zweigenbaum (LIMSI-CNRS, France)

Invited Speaker:

Chris Callison-Burch, University of Pennsylvania, US

Introduction to BUCC 2014

Comparable corpora are collections of documents that are comparable in content and form in various degrees and dimensions. This definition includes many types of parallel and non-parallel multilingual corpora, but also sets of monolingual corpora that are used for comparative purposes. Research on comparable corpora is active but used to be scattered among many workshops and conferences. The workshop series on "Building and Using Comparable Corpora" (BUCC) aims at promoting progress in this exciting emerging field by bundling its research, thereby making it more visible and giving it a better platform.

Following the six previous editions of the workshop which took place in Africa (LREC'08 in Marrakech), America (ACL'11 in Portland), Asia (ACL-IJCNLP'09 in Singapore), Europe (LREC'10 in Malta and ACL'13 in Sofia) and also on the border between Asia and Europe (LREC'12 in Istanbul), the workshop this year is co-located with LREC'14 in the middle of the Atlantic in Reykjavík, Iceland. The main theme for the current edition is "Comparable Corpora and Machine Translation". This topic reminds of the very origin of research in comparable corpora, which stemmed from the scarcity of parallel resources for Machine Translation (and also for Term Alignment).

We would like to thank all people who in one way or another helped in making this workshop once again a success. Our special thanks go to Chris Callison-Burch for accepting to give the invited presentation, to the members of the program committee who did an excellent job in reviewing the submitted papers under strict time constraints, and to the LREC'14 workshop chairs and organizers. Last but not least we would like to thank our authors and the participants of the workshop.

Pierre Zweigenbaum, Serge Sharoff, Reinhard Rapp, Ahmet Aker, Stephan Vogel

Crowdsourcing Translation
Chris Callison-Burch
Modern approaches to machine translation are data-driven. Statistical translation models are trained using parallel text, which consist of sentences in one language paired with their translation into another language. One advantage of statistical translation models is that they are language independent, meaning that they can be applied to any language that we have training data for. Unfortunately, most of the world's languages do not have sufficient amounts of training data to achieve reasonable translation quality.

In this talk, I will detail my experiments using Amazon Mechanical Turk to create crowd-sourced translations for "low resource" languages that we do not have training data for. I will discuss the following topics:

- Quality control: Can non-expert translators produce translations approaching the level of professional translators?
- Cost: How much do crowdsourced translations cost compared to professional translations?
- Impact of quality on training: When training a statistical model, what is the appropriate trade-off between small amounts of high quality data v. larger amounts of lower quality data?
- Languages: Which low resource languages is it possible to translate on Mechanical Turk? What volumes of data can we collect, and how fast?
- Implications: What implications does this have for national defense, disaster response, computational linguistics research, and companies like Google?

Construction of a French-LSF corpus
Michael Filhol and Xavier Tannier
In this article, we present the first academic comparable corpus involving written French and French Sign Language. After explaining our initial motivation to build a parallel set of such data, especially in the context of our work on Sign Language modelling and our prospect of machine translation into Sign Language, we present the main problems posed when mixing language channels and modalities (oral, written, signed), discussing the translation-vs-interpretation narrative in particular. We describe the process followed to guarantee feature coverage and exploitable results despite a serious cost limitation, the data being collected from professional translations. We conclude with a few uses and prospects of the corpus.

Merging Comparable Data Sources for the Discrimination of Similar Languages: The DSL Corpus Collection
Marcos Zampieri, Nikola Ljubesic and Jorg Tiedemann
This paper presents the compilation of the DSL corpus collection created for the DSL (Discriminating Similar Languages) shared task to be held at the VarDial workshop at COLING 2014. The DSL corpus collection were merged from three comparable corpora to provide a suitable dataset for automatic classification to discriminate similar languages and language varieties. Along with the description of the DSL corpus collection we also present results of baseline discrimination experiments reporting performance of up to 87.4% accuracy.

Building comparable corpora from social networks
Marwa Trabelsi, Malek Hajjem and Chiraz Latiri
Working with comparable corpora has proven an interesting alternative to rare parallel corpora in different Natural language tasks. Therefore many researchers have accentuated the need for large quantities of such corpora and the duty of works on their construction. In this paper, we highlight the interest and usefulness of textual data mining in social networks. We propose the exploitation of tweets from the microblog Twitter in order to construct comparable corpora. This work aims to develop a new method for the construction of comparable corpora that could be used later in multilingual information retrieval (MLIR), in Statistical Machine Translation (SMT) and in other fields.

Twitter as a Comparable Corpus to build Multilingual Affective Lexicons
Amel Fraisse and Patrick Paroubek
The main issue of any lexicon-based sentiment analysis system is the lack of affective lexicon. Such lexicons contain lists of words annotated with their affective classes. There exist some number of such resources but only for a few number of language and affective classes are, generally, reduced to two classes (positive and negative). In this paper we propose to use Twitter as a comparable corpus to generate a fine-grained and multilingual affective lexicons. Our approach is based in the co-occurence between English and target affective words in the same emotional corpus. And it can be applied for any target language. We experiment it to generate affective lexicons for seven languages (en, fr, de, it, es, pt, ru).

Multimodal Comparable Corpora for Machine Translation
Haithem Afli, Loïc Barrault and Holger Schwenk
The construction of a statistical machine translation (SMT) requires parallel corpus for training the translation model and monolingual data to build the target language model. A parallel corpus, also called bitext, consists in bilingual/multilingual texts. Unfortunately, parallel texts are a sparse resource for many language pairs. One way to overcome this lack of data is to exploit comparable corpora which are much more easily available. In this paper, we present the corpus developed for automatic parallel data extraction from multimodal comparable corpora, from Euronews and TED web sites. We describe the content of each corpus and how we extracted the parallel data with our new extraction system. We present our methods developed for multimodal corpora exploitation and discuss results on bitexts extracted.

Extended Translation Memories for Multilingual Document Authoring
Jean-Luc Meunier and Marc Dymetman
This paper proposes a small set of extensions to be made on a translation memory to support multilingual authoring. We describe how an instance of such extended formalism can be conveniently created thanks to a domain specific language. We also describe a motivating business use and describe how we implemented a full system. Finally, we report on the experiment we ran in a real business setting.

Using partly multilingual patents to support research on multilingual IR by building translation memories and MT systems
Lingxiao Wang, Christian Boitet and Mathieu Mangeot
In this paper, we describe the extraction of directional translation memories (TMs)from a partly multilingual corpus of patent documents, namely the CLEF-IP collection and the subsequent production and gradual improvement of MT systems for the associated sublanguages (one for each language), the motivation being to support the work of researchers of the MUMIA community. First, we analysed the structure of patent documents in this collection, and extracted multilingual parallel segments (English-German, English-French, and French-German) from it, taking care to identify the source language, as well as monolingual

segments. Then we used the extracted TMs to construit statistical machine translation systems (SMT). In order to get more parallel segments, we also imported monolingual segments into our post-editing system, and post-edited them with the help of SMT.

Comparability of Corpora in Human and Machine Translation
Ekaterina Lapshinova-Koltunski and Santanu Pal
In this study, we demonstrate a negative result from a work on comparable corpora which forces us to address a problem of comparability in both human and machine translation. We state that it is not always defined similarly, and corpora used in contrastive linguistics or human translation analysis cannot always be applied for statistical machine translation (SMT). So, we revise the definition of comparability and show that some notions from translatology, i.e. registerial features, should also be considered in machine translation (MT).

Identifying Japanese-Chinese Bilingual Synonymous Technical Terms from Patent Families
Zi Long, Lijuan Dong, Takehito Utsuro, Tomoharu Mitsuhashi and Mikio Yamamoto
In the task of acquiring Japanese-Chinese technical term translation equivalent pairs from parallel patent documents, this paper considers situations where a technical term is observed in many parallel patent sentences and is translated into many translation equivalents and studies the issue of identifying synonymous translation equivalent pairs. First, we collect candidates of synonymous translation equivalent pairs from parallel patent sentences. Then, we apply the Support Vector Machines (SVMs) to the task of identifying bilingual synonymous technical terms, and achieve the performance of over 85% precision and over 60% F-measure. We further examine two types of segmentation of Chinese sentences, i.e., by characters and by morphemes, and integrate those two types of segmentation in the form of the intersection of SVM judgments, which achieved over 90% precision.

Revisiting comparable corpora in connected space
Pierre Zweigenbaum
Bilingual lexicon extraction from comparable corpora is generally addressed through two monolingual distributional spaces of context vectors connected through a (partial) bilingual lexicon. We sketch here an abstract view of the task where these two spaces are embedded into one common bilingual space, and the two comparable corpora are merged into one bilingual corpus. We show how this paradigm accounts for a variety of models proposed so far, and where a set of topics addressed so far take place in this framework: degree of comparability, ambiguity in the bilingual lexicon, where parallel corpora stand with respect to this view, e.g., to replace the bilingual lexicon. A first experiment, using comparable corpora built from parallel corpora, illustrates one way to put this framework into practice. We also outline how this paradigm suggests directions for future investigations. We finally discuss the current limitations of the model and directions to solve them.

Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools

27 May 2014

ABSTRACTS

Editors:

Hend S. Al-Khalifa and Abdulmohsen Al-Thubaity

Workshop Programme

Date 27 May 2014

09:00 – 10:30 Session 1

09:00 – 09:20 Welcome and Introduction by Workshop Chair

09:20 – 09:50 Key note Speech by Prof. Mansour Algamdi
King Abdullah Initiative for Arabic Content

09:50 – 10:10 Wajdi Zaghouani
Critical Survey of the Freely Available Arabic Corpora

10:10 -10:30 Jonathan Forsyth
Automatic Readability Prediction for Modern Standard Arabic

10:30 – 11:00 Coffee break

11:00 -13:00 Session 2

11:00 – 11:20 Eshrag Refaee and Verena Rieser
Subjectivity and Sentiment Analysis of Arabic Twitter Feeds with Limited Resources

11:20 – 11:40 Ali Meftah, Yousef Alotaibi and Sid-Ahmed Selouani
Designing, Building, and Analyzing an Arabic Speech Emotional Corpus

11:40 – 12:00 Thomas Eckart, Uwe Quasthoff, Faisal Alshargi and Dirk Goldhahn
Large Arabic Web Corpora of High Quality: The Dimensions Time and Origin

12:00 – 12:20 Ryan Cotterell and Chris Callison-Burch
An Algerian Arabic / French Code-Switched Corpus

12:20:12:40 Mourad Loukam, Amar Balla and Mohamed Tayeb Laskri
An Open Platform, Based on Hpsg Formalism, for the Standard Arabic Language

12:40:13:00 Ghania Droua-Hamdani, Yousef Alotaibi, Sid-Ahmed Selouani and Malika Boudraa
Rhythmic Features across Modern Standard Arabic and Arabic Dialects

Workshop Organizers

Hend S. Al-Khalifa	King Saud University, KSA
Abdulmohsen Al-Thubaity	King AbdulaAziz City for Science and Technology, KSA

Workshop Programme Committee

Eric Atwell	University of Leeds, UK
Khaled Shaalan	The British University in Dubai (BUiD), UAE
Dilworth Parkinson	Brigham Young University, USA
Nizar Habash	Columbia University, USA
Khurshid Ahmad	Trinity College Dublin, Ireland
Abdulmalik AlSalman	King Saud University, KSA
Maha Alrabiah	King Saud University, KSA
Saleh Alosaimi	Imam University, KSA
Sultan almujaiwel	King Saud University , KSA
Adam Kilgarriff	Lexical Computing Ltd, UK
Amal AlSaif	Imam University, KSA
Maha AlYahya	King Saud University ,KSA
Auhood AlFaries	King Saud University , KSA
Salwa Hamadah	Taibah University, KSA
Abdullah Alfaifi	University of Leeds, UK

Preface

For Natural Language Processing (NLP) and Computational Linguistics (CL) communities, it was a known situation that Arabic is a resource poor language. This situation was thought to be the reason why there is a lack of corpus based studies in Arabic. However, the last years witnessed the emergence of new considerably free Arabic corpora and in lesser extent Arabic corpora processing tools.

This workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools (OSACT) aimed to encourage researchers and developers to foster the utilization of freely available Arabic corpora and open source Arabic corpora processing tools and help in highlighting the drawbacks of these resources and discuss techniques and approaches on how to improve them.

OSACT had an acceptance rate of 67%, we received 12 papers from which 8 papers were accepted. We believe the accepted papers are high quality and present mixture of interesting topics. We would like to thank all people who in one way or another helped in making this workshop a success. Our special thanks go to Professor Mansour Alghamdi for accepting to give the invited presentation, to the members of the program committee who did an excellent job in reviewing the submitted papers, to Saad Alotaibi for designing and updating OSACT website and to the LREC organizers. Last but not least we would like to thank our authors and the participants of the workshop.

Hend Al-Khalifa and Abdulmohsen Al-Thubaity
Reykjavik (Iceland), 2014

Session 1

Tuesday 27 May, 9:00 -10:30
Chairperson: Abdulmohsen Al-Thubaity

Keynote Speech: King Abdullah Initiative for Arabic Content
Prof. Mansour Algamdi

Critical Survey of the Freely Available Arabic Corpora

Wajdi Zaghouani

Abstract
The availability of corpora is a major factor in building natural language processing applications. However, the costs of acquiring corpora can prevent some researchers from going further in their endeavours. The ease of access to freely available corpora is urgent needed in the NLP research community especially for language such as Arabic. Currently, there is not easy was to access to a comprehensive and updated list of freely available Arabic corpora. We present in this paper, the results of a recent survey conducted to identify the list of the freely available Arabic corpora and language resources. Our preliminary results showed an initial list of 66 sources. We presents our findings in the various categories studied and we provided the direct links to get the data when possible.

Automatic Readability Prediction for Modern Standard Arabic

Jonathan Forsyth

Abstract
Research for automatic readability prediction of text has increased in the last decade and has shown that various machine learning (ML) methods can effectively address this problem. Many researchers have applied ML to readability prediction for English, while Modern Standard Arabic (MSA) has received little attention. Here I describe a system which leverages ML to automatically predict the readability of MSA. I gathered a corpus comprising 179 documents that were annotated with the Interagency Language Roundtable (ILR) levels. Then, I extracted lexical and discourse features from each document. Finally, I applied the Tilburg Memory-Based Learning (TiMBL) program to read these features and predict the ILR level of each document using 10-fold cross validation for both 3-way and 5-way classification tasks. I measured performance using the F-score. For 3-way and 5-way classifications my system achieved F-scores of 0.719 and 0.519 respectively. I discuss the implication of these results and the possibility of future development.

Session 2:

Tuesday 27 May, 11:00 -13:00
Chairperson: Mansour Algamdi

Subjectivity and Sentiment Analysis of Arabic Twitter Feeds with Limited Resources

Eshrag Refaee and Verena Rieser

Abstract
This paper addresses the task of automatic Subjectivity and Sentiment Analysis (SSA) for Arabic tweets. This is a challenging task because, first, there are no freely available annotated corpora available for this task, and second, most natural language processing (NLP) tools for Arabic are developed for Modern Standard Arabic (MSA) only and fail to capture the wide range of dialects

used in Arabic micro-blogs. In the following paper we show that, despite these challenges, we are able to learn a SSA classifier from limited amounts of manually annotated data, which reaches performance levels of up to 87.7% accuracy using cross-validation. However, an evaluation on a independent test set shows that these static models do not transfer well to new data sets, collected at a later point in time. An error analysis confirms that this drop in performance is due to topic-shifts in the twitter stream. Our next step is to extend our current models to perform semi-supervised online learning in order to continuously adapt to the dynamic nature of online media.

Designing, Building, and Analyzing an Arabic Speech Emotional Corpus

Ali Meftah, Yousef Alotaibi, Sid-Ahmed Selouani

Abstract
In this paper we describe a new emotional speech corpus recorded for Modern Standard Arabic (MSA). The newly designed corpus contains five target emotions, namely neutral, sadness, happy, surprised, and questioning, and it consists of 16 selected sentences that were read by 20 male and female native Arabic speakers. A human perceptual test was then applied to the recorded corpus. The test was performed by nine additional individuals. The results of the perceptual test verified the correctness of the intended emotions at a rate of 82.13%. Specifically, the most accurately identified emotion was "questioning," while the least identified emotion was "happy." Subsequent analyses of the results revealed that content sentences play an important role in influencing speakers with respect to controlling the intended emotion. This corpus is the first MSA corpus to be built using multiple variables, including gender, language content, and other speaker demographic data.

Large Arabic Web Corpora of High Quality: The Dimensions Time and Origin

Thomas Eckart, Faisal Alshargi, Uwe Quasthoff, Dirk Goldhahn

Abstract
Large textual resources are the basis for a variety of applications in the field of corpus linguistic. For most languages spoken by large user groups a comprehensive set of these corpora are constantly generated and exploited. Unfortunately for modern Arabic there are still shortcomings that interfere with systematic text analysis. The use of the Arabic language in many countries with different cultural backgrounds and the political changes in many of these countries over the last years require a broad and steady text acquisition strategy to form a basis for extended analysis. This paper describes the Arabic part of the Leipzig Corpora Collection (LCC) which is a provider of freely available resources for more than 200 languages. The LCC focuses on providing modern text corpora and wordlists via web-based interfaces for the academic community. As an example for the exploitation of these resources it will be shown how wordlists reflect political and cultural concepts that can be automatically exploited for diachronic or spatial comparisons.

An Algerian Arabic-French Code-Switched Corpus

Ryan Cotterell, Adithya Renduchintala, Naomi Saphra, Chris Callison-Burch

Abstract
Arabic is not just one language, but rather a collection of dialects in addition to Modern Standard Arabic (MSA). While MSA is used in formal situations, dialects are the language of everyday life. Until recently, there was very little dialectal Arabic in written form. With the advent of social-media, however, the landscape has changed. We provide the first romanized code-switched Algerian Arabic-French corpus annotated for word-level language id. We review the history and

sociological factors that make the linguistic situation in Algerian unique and highlight the value of this corpus to the natural language processing and linguistics communities. To build this corpus, we crawled an Algerian newspaper and extracted the comments from the news story. We discuss the informal nature of the language in the corpus and the challenges it will present. Additionally, we provide a preliminary analysis of the corpus. We then discuss some potential uses of our corpus of interest to the computational linguistics community.

An Open Platform Based on HPSG Formalism for the Standard Arabic Language

Mourad Loukam, Amar Balla and Mohamed Tayeb Laskri

Abstract
HPSG formalism knows since many years a great development in NLP. We are working on the HPSG formalism on the double aspect of modelling / implementation with the aim of its application to the standard Arabic language. In this paper, we present an open platform, based on HPSG formalism, for the standard Arabic language. The originality of the platform is that it is an integrated tool which offers the complete chain of parsing texts in Arabic language in order to produce their HPSG analysis. In the medium-term, our objective is to use the platform for developing applications for Arabic NLP.

Rhythmic Features across Modern Standard Arabic and Arabic Dialects

Ghania Droua-Hamdani, Yousef A. Alotaibi, Sid-Ahmed Selouani, Malika Boudraa

Abstract
This paper describes the speech timing of Modern Standard Arabic (MSA) using rhythmic features. Two main approaches are used to compute rhythm metrics: Interval Measures and Pairwise variability indices. The first approach consists of a comparison of MSA rhythm metrics computed from the ALGerian Arabic Speech Database with those resulting from using the West Point Arabic Speech corpus. The second approach compares these results with the rhythm measurements of six Arabic dialects. Many unexpected results are observed concerning rhythm variation within MSA and the dialects according to speakers' localities.

WILDRE2 - 2nd Workshop on Indian Language Data: Resources and Evaluation

27 May 2014

ABSTRACTS

Editors:

Girish Nath Jha, Kalika Bali, Sobha L, Esha Banerjee

Workshop Programme

27th May 2014

14.00 – 15.15 hrs: Inaugural session

14.00 – 14.10 hrs – Welcome by Workshop Chairs
14.10 – 14.30 hrs – Inaugural Address by Mrs. Swaran Lata, Head, TDIL, Dept of IT, Govt of India
14.30 – 15.15 hrs – Keynote Lecture by Prof. Dr. Dafydd Gibbon, Universität Bielefeld, Germany

15.15 – 16.00 hrs – Paper Session I
Chairperson: **Zygmunt Vetulani**

- Sobha Lalitha Devi, Vijay Sundar Ram and Pattabhi RK Rao, *Anaphora Resolution System for Indian Languages*
- Sobha Lalitha Devi, Sindhuja Gopalan and Lakshmi S, *Automatic Identification of Discourse Relations in Indian Languages*
- Srishti Singh and Esha Banerjee, *Annotating Bhojpuri Corpus using BIS Scheme*

16.00 – 16.30 hrs – Coffee break + Poster/Demo Session
Chairperson: **Kalika Bali**

- Niladri Sekhar Dash, *Developing Some Interactive Tools for Web-Based Access of the Digital Bengali Prose Text Corpus*
- Krishna Maya Manger, *Divergences in Machine Translation with reference to the Hindi and Nepali language pair*
- András Kornai and Pushpak Bhattacharyya, *Indian Subcontinent Language Vitalization*
- Niladri Sekhar Dash, *Generation of a Digital Dialect Corpus (DDC): Some Empirical Observations and Theoretical Postulations*
- S Rajendran and Arulmozi Selvaraj, *Augmenting Dravidian WordNet with Context*
- Menaka Sankarlingam, Malarkodi C S and Sobha Lalitha Devi, *A Deep Study on Causal Relations and its Automatic Identification in Tamil*
- Panchanan Mohanty, Ramesh C. Malik & Bhimasena Bhol, *Issues in the Creation of Synsets in Odia: A Report*
- Uwe Quasthoff, Ritwik Mitra, Sunny Mitra, Thomas Eckart, Dirk Goldhahn, Pawan Goyal, Animesh Mukherjee, *Large Web Corpora of High Quality for Indian Languages*
- Massimo Moneglia, Susan W. Brown, Aniruddha Kar, Anand Kumar, Atul Kumar Ojha, Heliana Mello, Niharika, Girish Nath Jha, Bhaskar Ray, Annu Sharma, *Mapping Indian Languages onto the IMAGACT Visual Ontology of Action*
- Pinkey Nainwani, *Handling Conflational Divergence in a pair of languages: the case of English and Sindhi*
- Jayendra Rakesh Yeka, Vishnu S G and Dipti Misra Sharma, *Semi automated annotated treebank construction for Hindi and Urdu*
- Saikrishna Srirampur, Deepak Kumar Malladi and Radhika Mamidi, *Improvised and Adaptable Statistical Morph Analyzer (SMA++)*

- K Kabi Khanganba and Girish Nath Jha, *Challenges in Indian Language Transliteration: a case of Devanagari, Bangla and Manipuri*
- Girish Nath Jha, Lars Hellan, Dorothee Beermann, Srishti Singh, Pitambar Behera and Esha Banerjee, *Indian languages on the TypeCraft platform – the case of Hindi and Odia*

16.30 – 17.30 hrs – Paper Session II
Chairperson: **Dr. S. S. Aggarwal**

- Nripendra Pathak and Girish Nath Jha, *Issues in Mapping of Sanskrit-Hindi Verb forms*
- Atul Kr. Ojha, Akanksha Bansal, Sumedh Hadke and Girish Nath Jha, *Evaluation of Hindi-English MT Systems*
- Sreelekha S, Pushpak Bhattacharyya and Malathi D, *Lexical Resources for Hindi Marathi MT*
- Esha Banerjee, Akanksha Bansal and Girish Nath Jha, *Issues in chunking parallel corpora: mapping Hindi-English verb group in ILCI*

17:30 – 18.10 hrs – Panel discussion *India and Europe - making a common cause in LTRs* Coordinator: **Hans Uszkoreit**

Panelists - Joseph Mariani, Swaran Lata, Zygmunt Vetulani, Dafydd Gibbon, Panchanan Mohanty

18:10- 18:25 hrs – Valedictory Address by Prof. Nicoletta Calzolari, CNR-ILC, Italy
18:25-18:30 hrs – Vote of Thanks

Workshop Organizers

Girish Nath Jha	Jawaharlal Nehru University, New Delhi
Kalika Bali	Microsoft Research Lab India, Bangalore
Sobha L	AU-KBC Research Centre, Anna University, Chennai

Workshop Programme Committee

A. Kumaran	Microsoft Research, India
Amba Kulkarni	University of Hyderabad, India
Chris Cieri, LDC	University of Pennsylvania
Dafydd Gibbon	Universität Bielefeld, Germany
Dipti Mishra Sharma	IIIT, Hyderabad, India
Girish Nath Jha	Jawaharlal Nehru University, New Delhi, India
Hans Uszkoreit	Saarland University, Germany
Indranil Datta	EFLU, Hyderabad, India
Jopseph Mariani	LIMSI-CNRS, France
Jyoti Pawar	Goa University, India
Kalika Bali	MSRI, Bangalore, India
Karunesh Arora	CDAC Noida, India
Malhar Kulkarni	IIT Bombay, India
Monojit Choudhary	Microsoft Research, India
Nicoletta Calzolari	CNR-ILC, Pisa, Italy
Niladri Shekhar Dash	ISI Kolkata, India
Panchanan Mohanty	University of Hyderabad, India
Pushpak Bhattacharya	IIT Bombay, India
S. S. Aggarwal	KIIT, Gurgaon
Sobha L	AU-KBC RC, Anna University, Chennai, India
Umamaheshwar Rao	University of Hyderabad, India
Zygmunt Vetulani	Adam Mickiewicz University, Poznan, Poland

Introduction

WILDRE – the 2nd workshop on Indian Language Data: Resources and Evaluation is being organized in Reykjavik, Iceland on 27th May, 2014 under the LREC platform. India has a huge linguistic diversity and has seen concerted efforts from the Indian government and industry towards developing language resources. European Language Resource Association (ELRA) and its associate organizations have been very active and successful in addressing the challenges and opportunities related to language resource creation and evaluation. It is therefore a great opportunity for resource creators of Indian languages to showcase their work on this platform and also to interact and learn from those involved in similar initiatives all over the world.

The broader objectives of the 2nd WILDRE will be

- to map the status of Indian Language Resources
- to investigate challenges related to creating and sharing various levels of language resources
- to promote a dialogue between language resource developers and users
- to provide opportunity for researchers from India to collaborate with researchers from other parts of the world

The call for papers received a good response from the Indian language technology community. Out of 29 full papers received for review, we selected 7 papers for oral, 13 for poster and 1 for demo presentation.

Paper Session I

Tuesday, 27 May, 15.15 – 16.00 hrs

Chairperson: Zygmunt Vetulani

Anaphora Resolution System for Indian Languages

Sobha Lalitha Devi, Vijay Sundar Ram R., and Pattabhi RK Rao

We describe our work on anaphora resolution engine for all Indian languages. Indian languages are morphologically rich. Morphological richness of the Indian languages is tapped to come up with generic anaphora resolution engine. The system architecture is designed in a simple plug-n-play model. We have used a machine learning technique, Conditional Random Fields (CRFs) to build the core engine. We have tested the engine with Tamil, a Dravidian language and Hindi, an Indo-Aryan language. The results are encouraging.

Automatic Identification of Discourse Relations in Indian Languages

Sobha Lalitha Devi, Sindhuja Gopalan, Lakshmi S

This paper describes the first effort on automatic identification of connectives and their arguments for three Indian languages Hindi, Malayalam and Tamil. We have adopted machine learning technique Conditional Random Fields (CRFs) for our work. We have used a corpus of 3000 sentences belonging to health domain. Domain independent features were extracted to improve the performance of the system. We mainly concentrated on the identification of explicit connectives and their arguments. Two sets of experiments were performed. First set of experiment was performed for the identification of connectives and next for the identification of argument boundaries. Using this approach we obtained encouraging results for all the three languages. Error analysis shows the presence of different structural patterns of discourse relations among three languages.

Annotating Bhojpuri Corpus using BIS Scheme

Srishti Singh and Esha Banerjee

The present paper talks about the application of the Bureau of Indian Standards (BIS) scheme for one of the most widely spoken Indian languages ‘Bhojpuri’. Bhojpuri has claimed for its inclusion in the Eighth Schedule of the Indian Constitution, where currently 22 major Indian languages are already enlisted. Recently through Indian government initiatives these scheduled languages have received the attention from Computational aspect, but unfortunately this non-scheduled language still lacks such attention for its development in the field of NLP. The present work is possibly the first of its kind.The BIS tagset is an Indian standard designed for tagging almost all the Indian languages. Annotated corpora in Bhojpuri and the simplified annotation guideline to this tagset will serve as an important tool for such well-known NLP tasks as POS- Tagger, Phrase Chunker, Parser, Structural Transfer, Word Sense Disambiguation (WSD), etc.

Poster Session

Tuesday, 27 May, 16.00 – 16.30 hrs

Chairperson: Kalika Bali

Indian Subcontinent Language Vitalization

András Kornai, Pushpak Bhattacharyya

We describe the planned Indian Subcontinent Language Vitalization (ISLV) project, which aims at turning as many languages and dialects of the subcontinent into digitally viable languages as feasible.

Augmenting Dravidian WordNet with Context

S. Rajendran, S. Arulmozi

It is difficult to interpret the meaning of a lexical item without context. WordNet lists different senses of a word and provides definition and usage example for each sense. But like any sense enumerative lexicon it also does not provide any mechanism for the novel usage of a word. The polysemy found in verbs and adjectives convincingly tell us that we have to augment WordNet with context. Such mechanism will help us to condense senses listed under a word and allow us to interpret the senses of a word creatively or generatively.

A Study on Causal Relations and its Automatic Identification in Tamil

Menaka S., Malarkodi C.S., Sobha Lalitha Devi

The objective of the present work is to identify the cause-effect expressions in Tamil text. We have classified the causal markers into different types for computational purposes. The linear order of cause-effect markers and arguments are explained with examples. Tamil corpora consisting of 31,741 sentences were annotated manually for this task. To overcome the structural interdependencies existing in cause-effect relations, we came up with the separate set of features for each type and the type specific models were generated. We have introduced the Sliding window type specific testing approach. Post-processing using linguistic and heuristic rules improves the system performance. We have conducted performance distribution and ten-fold cross validation experiments and the results are encouraging.

Issues in the Creation of Synsets in Odia: A Report

Panchanan Mohanty, Ramesh C. Malik & Bhimasena Bhol

Since languages differ from each other, it is difficult to find equivalents for the words and expressions of one language in another. So creating an interlingual WordNet in Odia vis-à-vis Hindi has been a challenging task. While dealing with the equivalence problems in Odia, creation of new expressions dominates the synsets involving various kinds of wage, derivation of nouns from nouns and adjectives, adjectives derived from nouns, and single-word Hindi synsets expressing complex ideas and kinship synsets. The other important procedure is borrowing that has been used widely in the domains of historical events, geographical locations, socio-cultural practices, place names, personal names, flora and

fauna, ecological entities, gods and goddesses, culture-specific items, etc.. Apart from these, certain problematic issues of the Hindi WordNet, viz. wrong categorization of synsets, concepts with inadequate information and description, mismatch between concepts and synsets, and imprecise concepts have also been discussed with a view to sensitizing other Indian language WordNet developers regarding these deficiencies. At the same time, we expect the quality of the Hindi WordNet to improve if these problems are taken care of.

Large Web Corpora of High Quality for Indian Languages

Uwe Quasthoff, Ritwik Mitra, Sunny Mitra, Thomas Eckart, Dirk Goldhahn, Pawan Goyal, Animesh Mukherjee

Large textual resources are the basis for a variety of applications in the field of corpus linguistics. For most languages spoken by large user groups a comprehensive set of these corpora are constantly generated and exploited. Unfortunately for modern Indian languages there are still shortcomings that interfere with systematic text analysis. This paper describes the Indian part of the Leipzig Corpora Collection which is a provider of freely available resources for more than 200 languages. This project focuses on providing modern text corpora and wordlists via web-based interfaces for the academic community. As an example for the exploitation of these resources it will be shown that they can be used for the visualization of semantic contexts of terms and for language comparison.

Mapping Indian Languages onto the IMAGACT Visual Ontology of Action

Massimo Moneglia, Susan Brown, Aniruddha Kar, Anand Kumar, Atul Kumar Ojha, Heliana Mello, Niharika, Girish Nath Jha, Bhaskar Ray and Annu Sharma.

Action verbs have many meanings, covering actions in different ontological types. Moreover, each language categorizes action in its own way. The range of variations within and across languages is largely unknown, causing trouble for natural language processing tasks and second language acquisition. IMAGACT is a corpus-based ontology of action concepts derived from English and Italian spontaneous speech resources, which makes use of the universal language of images to identify action types.IMAGACT4ALLis an internet infrastructure for mapping languages onto the ontology. Because the action concepts are represented with videos, extension into new languages is done using competence-based judgments by mother-tongue informants without intense lexicographic work involving underdetermined semantic description. It has been already proved on Spanish and Chinese and it is now in the process of being extended to Hindi, Bengali, Sanskrit and Portuguese. The paper presents the infrastructure and the methodology for mapping languages onto the ontology focussing on the features that make it a promising infrastructure for processing Indian languages.

Handling Conflational Divergence in a pair of languages: the case of English and Sindhi

Pinkey Nainwani

This paper discusses the nature of conflational divergence with reference to English and Sindhi. Dorr (1993) explained that a translation divergence arises when the natural translation of one language into another produce a very different form than that of the original. She demonstrated seven types of lexical-semantic divergences. One of them is conflational divergence which results when two or more words are required in one language to convey a

sense which is expressed by a single word in another language. Further, the paper describes the theoretical description of conflational divergence with reference to compound verbs, complex predicates, causative verbs, infinitival structures and many more. Due to the language complexities involved, I have adopted (S)tatistical (M)achine (T)ranslation approach to train English-Sindhi and Sindhi-English (parallel) corpora. In addition, it tries to illustrate to which extent SMT is able to capture sub-categorization of conflational divergence automatically.

Semi-automated annotated treebank construction for Hindi and Urdu

Jayendra Rakesh Yeka, Vishnu Ramagurumurthy, Dipti Misra Sharma

In this paper, we speak about the structure and paradigms chosen for creation of the annotated corpora for Hindi and Urdu. We briefly talk about the Shakti Standard Format that was chosen to suit needs of Indian language dependency annotation. This paper aims to present a framework for the creation of annotated corpus. We proceed to discuss the methods of automation chosen to overcome the laborious and time-consuming process of corpora annotation. We present the methods chosen to overcome the errors and multiple analyses that result through the task of annotation. We also present various methods used, both manual and automated, to ensure the quality of the treebank. We finally report the current status of the annotated corpora.

Improvised and Adaptable Statistical Morph Analyzer (SMA++)

Saikrishna Srirampur, Deepak Kumar Malladi, Radhika Mamidi

Morph analyzers play an important role in almost all the natural language applications. The morph analyzer (SMA++) we have developed is a data driven, statistical system. The system is a hybrid of the two best state of art statistical morph analyzers (SMA) viz. Morfette in Chrupała et al. (2008) and SMA in Malladi and Mannem (2013). We chose a robust feature set, which is a hybrid of the features used in the above SMAs. Our system predicts the gender, number, person, case (GNPC) and the lemma .The training and testing were done using the lib-linear classifier. Some rich features such as the morph tag of the previous token and the Part of Speech were used. Our goal is to come up with the best SMA which beats both the above SMAs. Our system is not language specific and can adapt to any language. Experimental results for the Indian language Hindi and sample results for Urdu have been shown, while results for other languages like Telugu etc. are in progress. The results for Hindi reflected higher accuracies than both of the above mentioned state of art SMAs.

Challenges in Indian Language Transliteration: a case of Devanagari, Bangla and Manipuri

K. Kabi Khanganba, Girish Nath Jha

The paper presents a Meitei-Bangla-Devanagari transliteration system and the challenges therein. Manipuri is a scheduled Indian language and was recently added to the Indian Language Corpora Initiative (ILCI) project being run in a consortium mode at Jawaharlal Nehru University. The Manipuri group faced difficulty in keying in Manipuri data in the

Meitei script as there are very few people who know typing in this script. Therefore, the project needed a transliteration system which could convert text written in Bengali script (which is known to most of the adult speaker) to Meitei. Automatic transliteration is a basic requirement in developing language technology in the diverse Indian language scenario. As most of the Indian scripts belong to the Brahmi family and have comparable sound systems, it is apparently not too difficult to create parallel arrays of utf charset encodings for comparing and substituting corresponding values between a pair of scripts. However, in reality they pose considerable challenges. Meitei presents special substitution challenges due to a slightly different representation scheme followed in Unicode for it. Another complication is due to the fact that in case of transliteration involving Meitei with another Indian language script (particularly from the north) we may be trying to substitute diverse phoneme sets leading to one-to-many and many-to-one matches.

Indian languages on the TypeCraft platform – the case of Hindi and Odia

Girish Nath Jha, Lars Hellan, Dorothee Beermann, Srishti Singh, Pitambar Behera, Esha Banerjee

This paper describes a plan for creating an aligned valence- and construction repository for Indian languages, starting with Hindi and Odia. The project will be a collaboration between the ILCI group at Jawaharlal Nehru University, India and the TypeCraft (TC) group at NTNU. In the first phase of the project, our task is to find a suitable theoretical framework for annotation at valence and construction level. In the second phase of the project (if the data download from the govt. of India data center site is opened to all), we will include a data portability and exchange module which will facilitate data import/export between the TC and the ILCI repositories.

Paper Session II

Tuesday, 27 May, 16.30 – 17.30 hrs

Chairperson: Dr. S. S. Aggarwal

Issues in Mapping of Sanskrit-Hindi Verb forms

Kumar Nripendra Pathak and Girish Nath Jha

Verb handling is the most important task for Machine Translation. A thorough syntactico-semantic study has been done in the Indian Grammatical Tradition which is highly appreciated by all the modern linguists worldwide. This paper deals with the syntactic patterns between Sanskrit-Hindi Verbs to formulate the possible algorithm to map the verbs for Sanskrit-Hindi Translator (SaHiT). This effort will help in producing linguistic rules for the required tool which can handle the verb forms in SaHiT.

Evaluation of Hindi-English MT Systems

Atul Kr. Ojha, Akanksha Bansal, Sumedh Hadke and Girish Nath Jha

Evaluation of any Machine Translation (MT) system is an important step towards improving its accuracy. In this paper, we are trying to evaluate Hindi-English module through two most

widely used MT systems - Bing (Microsoft) and Google. These MT systems are Statistics-Based MT systems (SBMT) and are capable of providing translation in many languages across the globe other than Hindi-English. For the purpose of evaluation, we tested Health and General cooking data and evaluated the English output text. Human evaluation strategy has been used for the purpose of evaluation, on the basis of which problem areas in both the MT systems were identified and compared to reach a conclusive analysis in terms of the output's fluency and comprehensibility. The comparative analysis helps in understanding not only which system is better but also what works best for automatic translation and under what circumstances. The discrepancies found are discussed with some suggestions towards their solution.

Lexical Resources for Hindi Marathi MT

Sreelekha. S, Pushpak Bhattacharyya, Malathi.D

In this paper we describe ways of utilizing lexical resources to improve the quality of statistical machine translation. We have augmented the training corpus with various lexical resources such as IndoWordnet semantic relation set, function words, kridanta pairs and verb phrases. We augmented parallel corpora in two ways (a) additional vocabulary and (b) inflected word forms. We have described case studies, evaluations and have given detailed error analysis for both Marathi to Hindi and Hindi to Marathi machine translation systems. From the evaluations we observed an order of magnitude improvement in translation quality. Lexical resources do help uplift performance when parallel corpora is scanty.

Issues in chunking parallel corpora: mapping Hindi-English verb group in ILCI

Esha Banerjee, Akanksha Bansal and Girish Nath Jha

A well annotated corpus is a treasure for Natural Language Processing (NLP) and can benefit NLP research activities like Machine Translation, Text Summarization and Information Retrieval. But since language is a dynamic and complex phenomenon, Part Of Speech (POS) annotation and Local Word Grouping or chunking prove to be challenging tasks mainly because of two reasons: first, maximum possible information about the structure of a sentence needs to be captured and second, the tags should be easy for the machine to map and facilitate desirable output resulting in an effective application. The present paper deals with issues faced in chunking verb groups in Hindi with respect to their mapping with English verb groups for machine translation. There are some verbal constructions in Hindi which are not present in English e.g. double causatives and serial constructions. Thus the task of mapping Hindi verbal groups with English for the purpose of translation can restrict the accuracy of the output attained. These divergences have been charted out with some relevant examples from both the languages. The purpose of describing these divergence issues is to find the most appropriate way of creating Chunk Annotation Tag-set standards which are currently under development for Indian languages.

9th SaLTMiL Workshop on
"Free/open-Source Language Resources for
the Machine Translation of Less-Resourced Languages"

27 May 2014

ABSTRACTS

Editors:

Mikel L. Forcada, Kepa Sarasola and Francis M. Tyers

Workshop Programme

09:00 – 09:30 Welcoming address by Workshop co-chair Mikel L. Forcada

09:30 – 10:30 Oral papers

Iñaki Alegria, Unai Cabezon, Unai Fernandez de Betoño, Gorka Labaka, Aingeru Mayor, Kepa Sarasola and Arkaitz Zubiaga
Wikipedia and Machine Translation: killing two birds with one stone

Gideon Kotzé and Friedel Wolff
Experiments with syllable-based English-Zulu alignment

10:30 – 11:00 Coffee break

11:00 – 13:00 Oral papers

Inari Listenmaa and Kaarel Kaljurand
Computational Estonian Grammar in Grammatical Framework

Matthew Marting and Kevin Unhammer
FST Trimming: Ending Dictionary Redundancy in Apertium

Hrvoje Peradin, Filip Petkovski and Francis Tyers
Shallow-transfer rule-based machine translation for the Western group of South Slavic languages

Alex Rudnick, Annette Rios Gonzales and Michael Gasser
Enhancing a Rule-Based MT System with Cross-Lingual WSD

13:00 – 13:30 General discussion

13:30 Closing

Workshop Organizers

Mikel L. Forcada	Universitat d'Alacant, Spain
Kepa Sarasola	Euskal Herriko Unibertsitatea, Spain
Francis M. Tyers	UiT Norgga árktalaš universitehta, Norway

Workshop Programme Committee

Iñaki Alegria	Euskal Herriko Unibertsitatea, Spain
Lars Borin	Göteborgs Universitet, Sweden
Elaine Uí Dhonnchadha	Trinity College Dublin, Ireland
Mikel L. Forcada	Universitat d'Alacant, Spain
Michael Gasser	Indiana University, USA
Måns Huldén	Helsingin Yliopisto, Finland
Krister Lindén	Helsingin Yliopisto, Finland
Nikola Ljubešić	Sveučilište u Zagrebu, Croatia
Lluís Padró	Universitat Politècnica de Catalunya, Spain
Juan Antonio Pérez-Ortiz	Universitat d'Alacant, Spain
Felipe Sánchez-Martínez	Universitat d'Alacant, Spain
Kepa Sarasola,	Euskal Herriko Unibertsitatea, Spain
Kevin P. Scannell	Saint Louis University, USA
Antonio Toral	Dublin City University, Ireland
Trond Trosterud	UiT Norgga árktalaš universitehta, Norway
Francis M. Tyers	UiT Norgga árktalaš universitehta, Norway

Introduction

The 9th International Workshop of the Special Interest Group on Speech and Language Technology for Minority Languages (SaLTMiL) will be held in Reykjavík, Iceland, on 27th May 2014, as part of the 2014 International Language Resources and Evaluation Conference (LREC). (For SALTMIL see: http://ixa2.si.ehu.es/saltmil/); it is also framed as one of the activities of European project Abu-Matran (http://www.abumatran.eu). Entitled "Free/open-source language resources for the machine translation of less-resourced languages", the workshop is intended to continue the series of SALTMIL/LREC workshops on computational language resources for minority languages, held in Granada (1998), Athens (2000), Las Palmas de Gran Canaria (2002), Lisbon (2004), Genoa (2006), Marrakech (2008), La Valetta (2010) and Istanbul (2012), and is also expected to attract the audience of Free Rule-Based Machine Translation workshops (2009, 2011, 2012).

The workshop aims to share information on language resources, tools and best practice, to save isolated researchers from starting from scratch when building machine translation for a less-resourced language. An important aspect will be the strengthening of the free/open-source language resources community, which can minimize duplication of effort and optimize development and adoption, in line with the LREC 2014 hot topic 'LRs in the Collaborative Age' (http://is.gd/LREChot).

Papers describe research and development in the following areas:

- Free/open-source language resources for rule-based machine translation (dictionaries, rule sets)
- Free/open-source language resources for statistical machine translation (corpora)
- Free/open-source tools to annotate, clean, preprocess, convert, etc. language resources for machine translation
- Machine translation as a tool for creating or enriching free/open-source language resources for less-resourced languages

Abstracts

Wikipedia and Machine Translation: killing two birds with one stone
Iñaki Alegria, Unai Cabezon, Unai Fernandez de Betoño, Gorka Labaka, Aingeru Mayor, Kepa Sarasola and Arkaitz Zubiaga

In this paper we present the free/open-source language resources for machine translation created in OpenMT-2 wikiproject, a collaboration framework that was tested with editors of Basque Wikipedia. Post-editing of Computer Science articles has been used to improve the output of a Spanish to Basque MT system called Matxin. The results show that this process can improve the accuracy of a Rule Based Machine Translation system in nearly 10% benefiting from the post-edition of 50,000 words in the Computer Science domain. We believe that our conclusions can be extended to MT engines involving other less-resourced languages lacking large parallel corpora or frequently updated lexical knowledge, as well as to other domains.

Experiments with syllable-based English-Zulu alignment
Gideon Kotzé and Friedel Wolff

As a morphologically complex language, Zulu has notable challenges aligning with English. One of the biggest concerns for statistical machine translation is the fact that the morphological complexity leads to a large number of words for which there exist very few examples in a corpus. To address the problem, we set about establishing an experimental baseline for lexical alignment by naively dividing the Zulu text into syllables, resembling its morphemes. A small quantitative as well as a more thorough qualitative evaluation suggests that our approach has merit, although certain issues remain. Although we have not yet determined the effect of this approach on machine translation, our first experiments suggest that an aligned parallel corpus with reasonable alignment accuracy can be created for a language pair, one of which is under-resourced, in as little as a few days. Furthermore, since very little language-specific knowledge was required for this task, our approach can almost certainly be applied to other language pairs and perhaps for other tasks as well.

Computational Estonian Grammar in Grammatical Framework
Inari Listenmaa and Kaarel Kaljurand

This paper introduces a new free and open-source linguistic resource for the Estonian language -- a computational description of the Estonian syntax and morphology implemented in Grammatical Framework (GF). Its main area of use is in controlled natural language applications, e.g. multilingual user interfaces to databases, but thanks to the recent work in robust parsing with GF grammars, is can also be used in wide-coverage parsing and machine translation applications together with other languages implemented as part of GF's Resource Grammar Library (RGL). In addition to syntax rules that implement all the RGL functions, this new resource includes a full paradigm morphological synthesizer for nouns, adjectives and verbs that works with 90%-100% accuracy depending on the number of input forms, as well as a general purpose monolingual lexicon of 80,000 words which was built from existing Estonian language resources.

FST Trimming: Ending Dictionary Redundancy in Apertium
Matthew Marting and Kevin Unhammer

The Free and Open Source rule-based machine translation platform Apertium uses Finite State Transducers (FST's) for analysis, where the output of the analyser is input to a second, bilingual FST. The bilingual FST is used to translate analysed tokens (lemmas and tags) from one language to another. We discuss certain problems that arise if the analyser contains entries that do not pass through the bilingual FST. In particular, in trying to avoid "half-translated" tokens, and avoid issues with the interaction between multiwords and tokenisation, language pair developers have created redundant copies of monolingual dictionaries, manually customised to fit their language pair. This redundancy gets in the way of sharing of data and bug fixes to dictionaries between language pairs. It also makes it more complicated to reuse dictionaries outside Apertium (e.g. in spell checkers). We introduce a new tool to trim the bad entries from the analyser (using the bilingual FST), creating a new analyser. The tool is made part of Apertium's lttoolbox package.

Shallow-transfer rule-based machine translation for the Western group of South Slavic languages
Hrvoje Peradin, Filip Petkovski and Francis Tyers

The South Slavic languages, spoken mostly in the Balkans, make up one of the three Slavic branches. The South Slavic branch is in turn comprised of two subgroups, the Eastern subgroup containing Macedonian and Bulgarian, and the western subgroup containing Serbo-Croatian and Slovenian. This paper describes the development of a bidirectional machine translation system for the western branch of South-Slavic languages — Serbo-Croatian and Slovenian. Both languages have a free word order, are highly inflected, and share a great degree of mutual inteligibility. They are also under-resourced as regards free/open-source resources. We give details on the resources and development methods used, as well as an evaluation, and general directions for future work.

Enhancing a Rule-Based MT System with Cross-Lingual WSD
Alex Rudnick, Annette Rios Gonzales and Michael Gasser

Lexical ambiguity is a significant problem facing rule-based machine translation systems, as many words have several possible translations in a given target language, each of which can be considered a sense of the word from the source language. The difficulty of resolving these ambiguities is mitigated for statistical machine translation systems for language pairs with large bilingual corpora, as large n-gram language models and phrase tables containing common multi-word expressions can encourage coherent word choices. For most language pairs these resources are not available, so a primarily rule-based approach becomes attractive. In cases where some training data is available, though, we can investigate hybrid RBMT and machine learning approaches, leveraging small and potentially growing bilingual corpora. In this paper we describe the integration of statistical cross-lingual word-sense disambiguation software with SQUOIA, an existing rule-based MT system for the Spanish-Quechua language pair, and show how it allows us to learn from the available bitext to make better lexical choices, with very few code changes to the base system. We also describe Chipa, the new open source CL-WSD software used for these experiments.

Legal Issues in Language Resources and Infrastructures

Workshop Programme

May 27, 2014

14.00 – Overview by the Organizing Committee

"What's happened in the last two years? EU, UK, US legal Developments"

Erik Ketzan, Institut für Deutsche Sprache / CLARIN-D

Prodromos Tsiavos, UCL / The Media Institute

Khalid Choukri, ELDA/ELRA

15:00 - Updates from language resources

Marc Kupietz, Institut für Deutsche Sprache

Denise DiPersio, Linguistic Data Consortium

Krister Linden, University of Helsinki, CLARIN ERIC

Andrejs Vasiljevs, Tilde, Latvia

Carla Parra, University of Bergen

Pavel Stranak, Charles University

16:45 – 17:15 Coffee break

17:15 – Updates from legal scholars

Pawel Kamocki, *Legal Aspects of Text and Data Mining*

Maarten Truyens, *Legal Aspects of Text Mining*

18:00 – Roundtable discussion

Workshop Organizers/Organizing Committee

Khalid Choukri	ELDA/ELRA
Erik Ketzan	Institut für Deutsche Sprache
Prodromos Tsiavos	UCL/ The Media Institute

Controlled Natural Language Simplifying Language Use

27 May 2014

ABSTRACTS

Editors:

Hitoshi Isahara, Key-Sun Choi, Shinhoi Lee, Sejin Nam

Workshop Programme

- Introduction

  09:00 – 09:25 Key-Sun Choi, Hitoshi Isahara
  "Workshop Introduction"

- Session 1 : CNL and Controlled Editing

  09:25 – 09:50 Pierrette Bouillon, Liliana Gaspar, Johanna Gerlach, Victoria Porro, Johann Roturier
  "Pre-editing by Forum Users: a Case Study"

  09:50 – 10:15 Wai Lok Tam, Yusuke Matsubara, Koiti Hasida, Motoyuki Takaai, Eiji Aramaki, Mai Miyabe, Hiroshi Uozaki
  "Generating Annotated Corpora with Autocompletion in a Controlled Language Environment"

- Session 2 : CNL Language Resource and Content Management

  10:15 – 10:30 Delyth Prys, David Chan, Dewi Bryn Jones
  "What is the Relationship between Controlled Natural Language and Language Registers?"

- Coffee Break

  10:30 – 11:00

- Invited Talk

  11:00 – 11:30 Teruko Mitamura (CMU, Language Technologies Institute)

- Session 2 Continued : CNL Language Resource and Content Management

  11:30 – 11:55 Kara Warburton
  "Developing Lexical Resources for Controlled Authoring Purposes"

  11:55 – 12:20 Giovanni Antico, Valeria Quochi, Monica Monachini, Maurizio Martinelli
  "Marrying Technical Writing with LRT"

- Session 3 : ISO Standard for CNL

  12:20 – 13:00 Key-Sun Choi, Hitoshi Isahara
  "Toward ISO Standard for Controlled Natural Language"

Workshop Organizers

Key-Sun Choi	KAIST
Hitoshi Isahara	Toyohashi University of Technology
Christian Galinski	Infoterm
Laurent Romary	INRIA

Workshop Programme Committee

Key-Sun Choi	KAIST
Hitoshi Isahara	Toyohashi University of Technology
Christian Galinski	Infoterm
Laurent Romary	INRIA

Preface

The study of controlled natural language has a long history due to its commercial impact as well as its effectiveness in applications like machine translation, librarianship, information management, terminology management, mobile communication, legal documents, and so on. On the other hand, “text simplification” is also beneficial for efficient communication with respect to all kinds of language use in the Web, such as simplified English Wikipedia for instance. The current progress of linked data also assumes a great potential for knowledge acquisition from text and web data, for example, NLP2RDF and its NIF (http://nlp2rdf.org). It is also obvious that its data fusion and knowledge fusion are more beneficial from the controlled or simplified text or structured source. There is also a working item on this topic in ISO/TC37 “Terminology and other language and content resources” for recommending the principles of controlled natural language and its supporting environment and utilization. This workshop is altogether to know about the scope of controlled natural language for simplifying use in the aspects of their pre-editing for controlled natural language use, their language resources and content management systems in technical writing and mobile life, and the interoperability and relation in the context of standardization. As a result, the workshop may identify the environment of controlled natural language use, their guideline to use, relationship with language resources and other systems, interlinking interoperability and dependency with other standards and activities, and discovery of controlled natural language for human technical writing as well as for the knowledge acquisition and knowledge fusion processes manually and/or automatically by computing and linking in the web environment. It is also observable to see the cooperative work items, to identify the shared tasks to work together open and to analyze their units for simplifying use of language.

Session 1: CNL and Controlled Editing

Tuesday 27 May, 09:25 – 10:15
Chairperson: Hitoshi Isahara

Pre-editing by Forum Users: a Case Study

Pierrette Bouillon, Liliana Gaspar, Johanna Gerlach, Victoria Porro, Johann Roturier

Previous studies have shown that pre-editing techniques can handle the extreme variability and uneven quality of user-generated content (UGC), improve its machine-translatability and reduce post-editing time. Nevertheless, it seems important to find out whether real users of online communities, which is the real life scenario targeted by the ACCEPT project, are linguistically competent and willing to pre-edit their texts according to specific pre-editing rules. We report the findings from a user study with real French-speaking forum users who were asked to apply pre-editing rules to forum posts using a specific forum plugin. We analyse the interaction of users with pre-editing rules and evaluate the impact of the users' pre-edited versions on translation, as the ultimate goal of the ACCEPT project is to facilitate sharing of knowledge between different language communities.

Generating Annotated Corpora with Autocompletion in a Controlled Language Environment

Wai Lok Tam, Yusuke Matsubara, Koiti Hasida, Motoyuki Takaai, Eiji Aramaki, Mai Miyabe, Hiroshi Uozaki

This paper presents a novel attempt to generate annotated corpora by making use of grammar based autocompletion. The idea is to automatically generate corpora on the fly while a user is working on his own stuff. In the medical domain, this user would be a physician. While he uses an authoring tool to write a pathology report or enters text in an Electronic Healthcare Record (EHR) system, grammar and ontology-based autocompletion is built into such systems such that user input is limited to text parseable by the grammar used for autocompletion. As soon as the user finishes his work, the grammar used for autocompletion would be used for assigning syntactic structures and semantic representations to his input automatically. This way corpora can be generated by limiting annotation and grammar writing by a linguist to help building grammar and ontology-based autocompletion into a EHR system or an authoring tool of pathology reports. After autocompletion starts working hand-in-hand with these applications, new input from users does not need further annotation by human. Users are not supposed to be paid for using an EHR system or an authoring tool with built-in autocompletion that helps them to do their job.

Session 2: CNL Language Resource and Content Management

Tuesday 27 May, 10:15 – 12:20 (Coffee Break and Invited Talk in-between)
Chairperson: Koiti Hasida

Developing Lexical Resources for Controlled Authoring Purposes

Kara Warburton

Controlled authoring is increasingly being adopted by global organisations as a means to improve their content and manage translation costs. In recent years, some controlled authoring software applications have become available. These applications assist writers in adhering to rules of style, grammar, and terminology. In this paper, the author describes her experiences in implementing lexical resources in controlled authoring software. Incorporating pre-existing terminological resources presents challenges because they were typically developed as an aid for translators. While there are economies of scale to be achieved by addressing the needs of controlled authoring and

translation from one lexical resource, careful planning is required in order to account for different requirements.

Marrying Technical Writing with LRT

Giovanni Antico, Valeria Quochi, Monica Monachini, Maurizio Martinelli

In the last years the Technical Writer operational scenarios and the workflow sensibly changed; specifically, “free style” writing – or manual writing - has become outdated and technical writing is now much more concerned with structured management of content than in the past. Technical writing has become more demanding due to a number of factors among which the rise and spread of mobile devices usage. This paper discusses the new needs of technical writing and content management business and how LRT can help it improve quality and productivity.

What is the relationship between Controlled Natural Language and Language Registers?

Delyth Prys, David Chan and Dewi Bryn Jones

This paper will examine the relationship between language register and controlled natural language, using as a case study an analysis of a 30 million-word corpus collected from an online spelling and grammar checker with a wide variety of text types, including email drafts, Twitter and Facebook posts, blogs, student essays, journalistic articles, simplified writings optimised for clarity and technical translations. Different registers are identified based on linguistic characteristics, automated tagging of register in the corpus is described, and the feasibility of applying similar techniques to identify controlled languages is discussed.

Session 3: ISO Standard for CNL

Tuesday 27 May, 12:20 – 13:00
Chairperson: Key-Sun Choi

Toward ISO Standard for Controlled Natural Language

Key-Sun Choi, Hitoshi Isahara

This standard is the first part of the series of ISO standards that are targeted at controlled natural language (CNL) in written languages. It focuses on the basic concepts and general principles of CNL that apply to languages in general. It will cover properties of CNL and CNL classification scheme. The subsequent parts will, however, focus on the issues specific to particular viewpoint and/or applications, such as particular CNLs, CNL interfaces, implementation of CNLs, and evaluation techniques for CNL.

Semantic Processing of Legal Texts (SPLeT-2014)

27 May 2014

ABSTRACTS

Editors:

Enrico Francesconi, Simonetta Montemagni, Wim Peters,

Giulia Venturi, Adam Wyner

Workshop Programme

14:00-15:00 – Introductory session

14:00-14:15 – Introduction by Workshop Chairs

14:15-15:00 – Invited Talk
Sophia Ananiadou (University of Manchester, National Centre for Text Mining, UK)
Adapting text mining from biology to the legal domain: what do we need?

15:00-16:00 – Paper session

15:00 –15:20
Tommaso Agnoloni, Lorenzo Bacci, Maria Teresa Sagri
Legal keyword extraction and decision categorization: a case study on italian civil case law

15:20 –15:40
Frane Šarić, Bojana Dalbelo Bašić, Marie-Francine Moens, Jan Šnajder
Multi-label Classification of Croatian Legal Documents Using EuroVoc Thesaurus

15:40 –16:00
Frantisek Cvrček, Karel Pala, Pavel Rychlý
Behaviour of Collocations in the Language of Legal Subdomains

16:00 – 16:30 Coffee break

16:30-17:10 – Paper session

16:30–16:50
Radboud Winkels, Jochem Douw, Sara Veldhoen
State of the ART: an Argument Reconstruction Tool

16:50–17:10
Tommaso Agnoloni
Network Analysis of Italian Constitutional Case Law

17:10-18:30 – Panel session
Panel on "Designing, constructing and using Legal Language Resources"
Moderator: Simonetta Montemagni (Istituto di Linguistica Computazionale "Antonio Zampolli")
Panelists: Cristina Bosco (Università di Torino), Guillaume Jacquet (European Commission Joint Research Centre, JRC), Alessandro Mazzei (Università di Torino), Karel Pala (Masaryk University), Daniela Tiscornia (Istituto di Teoria e Tecniche dell'Informazione Giuridica), Giulia Venturi (Istituto di Linguistica Computazionale "Antonio Zampolli"), Vern Walker (Hofstra University School of Law Hempstead)

Workshop Organizers

Enrico Francesconi	Istituto di Teoria e Tecniche dell'Informazione Giuridica del CNR, Florence, Italy
Simonetta Montemagni	Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR, Pisa, Italy
Wim Peters	Natural Language Processing Research Group, University of Sheffield, UK
Giulia Venturi	Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR, Pisa, Italy
Adam Wyner	Department of Computing Science, University of Aberdeen, UK

Workshop Programme Committee

Kevin Ashley	University of Pittsburgh, USA
Mohammad Al-Asswad	Cornell University, USA
Anderson Bertoldi	Universidade do Vale do Rio dos Sinos, Brazil
Danièle Bourcier	Humboldt Universität, Berlin, Germany
Thomas Bruce	LII Cornell, USA
Pompeu Casanovas	Institut de Dret i Tecnologia, UAB, Barcelona, Spain
Jack Conrad	Thomson-Reuters, USA
Michael Curtotti	Australian National University, Australia
Matthias Grabmair	University of Pittsburgh, USA
Marie-Francine Moens	Katholieke Universiteit Leuven, Belgium
Thom Neale	Sunlight Foundation, USA
Karel Pala	Masaryk University, Brno, Czech Republic
Paulo Quaresma	Universidade de Évora, Portugal
Erich Schweighofer	Universität Wien, Rechtswissenschaftliche Fakultät, Wien, Austria
Rolf Schwitter	Macquarie University, Australia
Daniela Tiscornia	Istituto di Teoria e Tecniche dell'Informazione Giuridica of CNR, Florence, Italy
Tom van Engers	Leibniz Center for Law, University of Amsterdam, Netherlands
Vern R. Walker	Hofstra University School of Law, Hofstra University, USA
Radboud Winkels	Leibniz Center for Law, University of Amsterdam, Netherlands

Preface

Since 2008, the LREC conference has provided a stimulating environment for the Workshop on "Semantic Processing of Legal Texts" (SPLeT) focusing on the topics of Language Resources (LRs) and Human Language Technologies (HLTs) in the legal domain. The workshops have been a venue where researchers from the Computational Linguistics and Artificial Intelligence and Law communities meet, exchange information, compare perspectives, and share experiences and concerns on the topic of legal knowledge extraction and management, with particular emphasis on the semantic processing of legal texts. Along with the SPLeT workshops, there have been a number of workshops and tutorials focusing on different aspects of semantic processing of legal texts at conferences of the Artificial Intelligence and Law community (e.g. JURIX, ICAIL).

To continue this momentum and to advance research, the 5th edition of SPLeT has been organised in conjunction with LREC-2014. LREC provides a forum in which to report on applications of linguistic technologies to particular domains as well as a context where individuals from academia and industry can interact to discuss problems and opportunities, find new synergies, and promote initiatives for international cooperation. Thus, the workshop at LREC is expected to bring to the attention of the broader LR/HLT community the specific technical challenges posed by the semantic processing of legal texts and also share with the community the motivations and objectives which make it of interest to researchers in legal informatics.

The last few years have seen a growing body of research and practice in the field of AI & Law which addresses a range of topics: automated legal argumentation, semantic and cross-language legal IR, document classification, legal drafting, legal knowledge extraction, as well as the construction of legal ontologies and their application. In this context, it is of paramount importance to use NLP techniques and tools as well as linguistic resources supporting the process of knowledge extraction from legal texts.

New to this edition of the workshop and in line with LREC 2014 Special Highlight we organized a panel on the topic of "Legal Language Resources" with the final aim of constructing a map of legal language resources, enabling their reuse (in reproducing and evaluating experiments) and extension. The resources presented and discussed in the panel range from annotated corpora to lexicons, thesauri and ontologies for different languages.

We would like to thank all the authors for submitting their research and the members of the Program Committee for their careful reviews and useful suggestions to the authors. Thanks are also due to the panelists who contributed the first panel organized around the topic of legal language resources. We would also like to thank our invited speaker, Sophia Ananiadou, for her contribution. Last but not least, we would like to thank the LREC 2014 Organising Committee that made this workshop possible.

The Workshop Chairs

Enrico Francesconi
Simonetta Montemagni
Wim Peters
Giulia Venturi
Adam Wyner

Session 1
Tuesday 27 May, 15:00 – 16:00

Legal keyword extraction and decision categorization: a case study on Italian civil case law

Tommaso Agnoloni, Lorenzo Bacci, Maria Teresa Sagri

Abstract
In this paper we present an approach to keyword extraction and automatic categorization of Italian case law of first instance on civil matters. The approach complements classic NLP based analysis of texts with legal and domain features extraction. The study originated from an experimental activity promoted by the Ministry of Justice for automated semantic metadata attribution in case law deposit in the framework of the digitalization of civil trial in Italy.

Multi-label Classification of Croatian Legal Documents Using EuroVoc Thesaurus

Frane Šarić, Bojana Dalbelo Bašić, Marie-Francine Moens, Jan Šnajder

Abstract
The automatic indexing of legal documents can improve access to legislation. EuroVoc thesaurus has been used to index documents of the European Parliament as well as national legislative. A number of studies exists that address the task of automatic EuroVoc indexing. In this paper we describe the work on EuroVoc indexing of Croatian legislative documents. We focus on the machine learning aspect of the problem. First, we describe the manually indexed Croatian legislative documents collection, which we make freely available. Secondly, we describe the multi-label classification experiments on this collection. A challenge of EuroVoc indexing is class sparsity, and we discuss some strategies to address it. Our best model achieves 79.6% precision, 60.2% recall, and 68.6% F1-score.

Behaviour of Collocations in the Language of Legal Subdomains

Frantisek Cvrček, Karel Pala, Pavel Rychlý

Abstract
In the paper we examine the collocational behaviour of multiword expression in legal sublanguages, i.e. in texts of statutory law, texts of case laws of Supreme Courts and law textbooks. We show that the comparison of collocations coming from the individual types of legal texts provides quantifiable data, which contain information about terminological nature of the observed language expressions. From the observations we made it follows that the legal language of the primary regulations considerably differs from the sublanguage of the secondary regulations. The quantitative analysis of the Czech legal texts has convincingly shown that the corpus analysis working with relatively simple means indicates the high number of changes in the texts of law regulations. In this way the changes also show that the corpus analysis also reflects the problems in our society – too many and fast changes in the legal texts prevent lawyers from the correct handling the individual court cases. In the paper we also exploit the results of the project PES (Právní elektronický slovínk, Legal Electronic Dictionary).

Session 2
Tuesday 27 May, 16:30 – 17:10

State of the ART: an Argument Reconstruction Tool

Radboud Winkels, Jochem Douw, Sara Veldhoen

Abstract
This paper describes the outcomes of a series of experiments in automated support for users that try to find and analyse arguments in natural language texts in the context of the FP7 project IMPACT. Manual extraction of arguments is a on-trivial task and requires extensive training and expertise. We investigated several possibilities to support this process by using natural language processing (NLP), from classifying pieces of text as either argumentative or non-argumentative to clustering answers to policy green paper questions in the hope that these clusters would contain similar arguments. Results are diverse, but also show that we cannot come a long way without an extensive pre-tagged corpus.

Network Analysis of Italian Constitutional Case Law

Tommaso Agnoloni

Abstract
In this paper we report on an ongoing research on the application of network metrics to the corpus of Italian constitutional case law. The research was enabled by the recent release as open data of the complete datasets of the jurisprudential production of the Italian Constitutional Court. The datasets include the complete textual corpora of Court judgements together with a rich set of associated structured metadata. Using the standard unique identifiers for case law recommended by the EU Council and a recently developed jurisprudential reference extractor, adapted to constitutional case law texts, we were able to construct the graph of jurisprudential references of Italian constitutional decisions. On the resulting network, first metrics have been evaluated and further research activities are foreseen exploiting the richness of the datasets and their potential connections.

Language Technology Service Platforms: Synergies, Standards, Sharing

31 May 2014

PROGRAMME

Editor:

Nicoletta Calzolari

Workshop Draft Programme

09:00 – 10:00 – Presentations of the initiatives involved in the organisation

10:00 – 10:30 – Session
Yohei Murakami et al., *Ontology for Language Service Interoperability*

Asuncion Gomez Perez, *Linked Data for sharing, discovery and re-use of Language Resources at a Web scale*

10:30 – 11:00 Coffee break

11:00 – 13:00 – Session
Nancy Ide et al., *LAPPS/ISO Exchange Vocabulary for web service interoperability*

Steve Cassidy, *Alveo: Combining Tools and Resources in National Infrastructure to support Human Communication Research*

Steven Krauwer, *Collaboration platforms: the CLARIN vision*

Eric Nyberg and Di Wang, *Open Advancement*

Hans Uszkoreit, *TBA*

Andrejs Vasiļjevs, *MLi – towards European multilingual infrastructure for public and industry services*

13:00 – 14:30 Lunch break

14:30 – 16:00 – Session
Khalid Choukri, *Language Resources identification and citation: implications for a platform*

Nicoletta Calzolari, *Enabling sharing and replicability of research results: consequences for a platform*

Ed Hovy, *The Role of Language Technology and Standards in Modernizing Scholarly Publication: The FORCE11 Initiative*

Chris Cieri and Denise DiPersio, *Licensing for globally distributed services*

16:00 – 16:30 Coffee break

16:30 – 18:30 *– Concluding Discussion – Visions and strategies for the future of Language Infrastructures: towards Common Resolutions*

Discussants: All the Speakers … and Workshop participants

Workshop Organisers

Nicoletta Calzolari - ILC-CNR, Italy and ELRA, France
Khalid Choukri - ELRA, France
Christopher Cieri - LDC, USA
Tomaž Erjavec - Jožef Stefan Institute, Slovenia
Nancy Ide - Vassar College, USA
Toru Ishida - Kyoto University, Japan
Oleksandr Kolomiyets - KU Leuven, Belgium
Joseph Mariani - LIMSI-CNRS and IMMI, France
Yohei Murakami - Kyoto University, Japan
Satoshi Nakamura - Nara Institute of Science and Technology, Japan
Senja Pollak - Jožef Stefan Institute, Slovenia
James Pustejovsky - Brandeis University, USA
Georg Rehm - DFKI GmbH, Germany
Herman Stehouwer - Max Planck, Germany
Hans Uszkoreit - DFKI GmbH, Germany
Andrejs Vasiļjevs - Tilde, Latvia
Peter Wittenburg - Max Planck Institute for Psycholinguistics, The Netherlands

Introduction

Motivation and background

Increasingly, Human Language Technology (HLT) requires sophisticated infrastructures to support research, development, innovation, collaboration and deployment as services ready for production use. To address this need, several supporting infrastructures have been established over the past few years, and others are being built or planned.

The LREC 2014 Workshop on Language Technology Service Platforms brings together major infrastructural/coordination initiatives from all over the world. The overall goal is to explore means by which these and other infrastructure projects can best collaborate and interoperate in order to avoid duplication of effort, fragmentation, and above all, to facilitate the use and reuse of technologies and resources distributed throughout the world.

Focus

Web services are an increasingly common means to provide access to language technologies and resources. These services typically work in combination with repositories of language resources and workflow managers. This development brings with it its own set of issues in relation to collaboration and interoperability, including:

- interoperability of input to and output from language technologies deployed as web services;
- means to provide services for evaluation/replicability of results and iterative development;
- means to support multi-site collaborative work;
- licensing and cataloguing of language technologies and resources;
- sharing and access mechanisms to language technologies and resources;
- quality assessment and sustainability of language technologies and resources.

Aims

This workshop aims to foster discussion on these (and related) issues in order to arrive at a set of concrete plans for future collaboration and cooperation as well as immediate next steps.

General discussions will focus on the following questions: How can the various infrastructures collaborate, in both the near and long-term future? What are the steps needed in order to share both language technologies and resources? How can the projects and initiatives (including not only those involved in the workshop, but also others) join forces in order to eventually create a global infrastructure for Human Language Technologies?

The goal is to leave the workshop with a resolution that 1. lists all active infrastructure and platform initiatives, 2. describes the consensus of all initiatives involved in the workshop, 3. outlines the requirements for collaboration and 4. proposes solutions.

Researchers and technologists interested in platforms, services, sharing of language resources etc. are encouraged to participate in the workshop in order to make sure that their voice is heard. As described above, the consensus and outcome of the workshop will be put down in writing in a short resolution document meant to be used by the whole community for public relation and dissemination purposes, especially with regard to discussions with journalists, administrators, politicians and funding agencies.

Preliminary program plan

The first session will provide short introductions to the infrastructural/coordination initiatives involved in the organisation.

In order to outline some concrete next steps for the immediate future, there will be sessions devoted to surveying two to four currently implemented solutions to crucial problems, with an eye toward assessing and comparing the various solutions in order to determine immediate action items. These sessions will address topics such as:

- interoperability and the use of standards, for example, syntax and semantics used to exchange information between web services and/or technologies that may not have been developed at the same site (i.e., that do not necessarily utilize the same formats, categories, etc.)
- implemented means to provide evaluation/replicability and means to enable multi-site collaboration
- licensing for data and tools shared over networks and services.

Contribute to an overview of the Language Resources and Technologies landscape!

In order to facilitate the discussion we ask workshop participants to answer the following questions and to send their answers to the organisers (see Contact mail below) at the beginning of May. A summary of the responses will be provided at the workshop to inform and to focus the discussion.

(1) *Access* – How do you make information about your tools and/or resources available to the world? How and where do you find information on tools and resources you would like to use?

(2) Obstacles to Data and Technology Exchange – What do you see as the major obstacle(s) to the exchange of data between technologies?

(3) *Data or Technology Gaps* – Are there tools, technologies or resources that do not exist at this time that are required to answer your research or development questions?

(4) *Interoperability and Standards* – What syntax and semantics do you use to exchange information between web services and/or tools that may not have been developed at the same site (i.e., do not necessarily utilize the same formats, categories, etc.)?

(5) *Evaluation* – What have you implemented to provide evaluation/replicability?

(6) *Licensing* – How are you handling licensing for data shared over networks and services?

(7) *Collaboration* – How would you propose to promote collaboration among the various infrastructure projects located around the world?

We also welcome any additional comments or views that you wish to express.

We look forward to welcoming you in Reykjavik!

6th Workshop on the Representation and Processing of Sign Languages: Beyond the Manual Channel

Reykjavik, Iceland, 31 May 2014

ABSTRACTS

Editors:

Onno Crasborn, Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, Julie Hochgesang, Jette Kristoffersen, Johanna Mesch

Workshop Programme

09:00 – 10:30	Session A – Oral/Signed: *Linguistic Signals in the Face*
10:30 – 11:00	Coffee break
11:00 – 13:00	Session B - Posters: *Annotation Issues*
13:00 – 14:00	Lunch break
14:00 – 16:00	Session C - Posters: *Language Technology*
16:00 – 16:30	Coffee break
16:30 – 18:00	Session D – Oral/Signed: *Integration*

Workshop Organizers

Onno Crasborn	Radboud University, Nijmegen NL
Eleni Efthimiou	Institute for Language and Speech Processing, Athens GR
Evita Fotinea	Institute for Language and Speech Processing, Athens GR
Thomas Hanke	Institute of German Sign Language, University of Hamburg, Hamburg DE
Julie Hochgesang	Gallaudet University, Washington US
Jette Kristoffersen	Centre for Sign Language, University College Capital, Copenhagen DK
Johanna Mesch	Stockholm University, Stockholm SE

Workshop Programme Committee

Richard Bowden	University of Surrey, Guildford GB
Penny Boyes Braem	Center for Sign Language Research, Basel CH
Annelies Braffort	LIMSI/CNRS, Orsay FR
Christophe Collet	IRIT, University of Toulouse, Toulouse FR
Kearsy Cormier	Deafness Cognition and Language Research Centre, London GB
Onno Crasborn	Radboud University, Nijmegen NL
Svetlana Dachkovsky	University of Haifa, Haifa IL
Eleni Efthimiou	Institute for Language and Speech Processing, Athens GR
Stavroula-Evita Fotinea	Institute for Language and Speech Processing, Athens GR
John Glauert	University of East Anglia, Norwich GB
Thomas Hanke	Institute of German Sign Language, University of Hamburg, Hamburg DE
Alexis Heloir	German Research Centre for Artificial Intelligence, Saarbrücken DE
Jens Heßmann	University of Applied Sciences Magdeburg-Stendal, Magdeburg DE
Julie Hochgesang	Gallaudet University, Washington US
Trevor Johnston	Macquarie University, Sydney AU
Reiner Konrad	Institute of German Sign Language, University of Hamburg, Hamburg DE
Jette Kristoffersen	Centre for Sign Language, University College Capital, Copenhagen DK
Lorraine Leeson	Trinity College, Dublin IE
Petros Maragos	National Technical University of Athens, Athens GR
John McDonald	DePaul University, Chicago US
Johanna Mesch	Stockholm University, Stockholm SE
Carol Neidle	Boston University, Boston US
Christian Rathmann	Institute of German Sign Language, University of Hamburg, Hamburg DE
Adam Schembri	National Institute for Deaf Studies and Sign Language, La Trobe University, Melbourne AU
Rosalee Wolfe	DePaul University, Chicago US

Session A: Linguistic Signals in the Face

Saturday 31 May, 09:00 – 10:30
Chairperson: Julie Hochgesang Oral/Signed Session

Discourse-based annotation of relative clause constructions in Turkish Sign Language (TID): A case study

Okan Kubus

The functions of relative clause constructions (RCC) should be ideally analyzed at the discourse level, since the occurrence of RCCs can be explained by looking at interlocutors' use of grammatical and intonational means (cf. Fox and Thompson, 1990). To date, RCCs in sign language have been analyzed at the syntactic level with a special focus on cross-linguistic comparisons (see e.g. Pfau and Steinbach, 2005; Branchini and Donati, 2009). However, to our knowledge, there is no systematic corpus-based analysis of RCCs in sign languages so far. Since the elements of RCCs are mostly non-manual markers, it is often unclear how to capture and tag these elements together with the functions of RCCs. This question is discussed in light of corpus-based data from Turkish Sign Language. Following Biber et al. (2007), the corpus-based analysis of RCCs in TID follows the "top-down" approach. In spite of modality-specific issues, the steps in the process of annotation and identification of RCCs in TID fairly resemble this approach. The advantage of using these multiple steps is that the procedure not only captures the discourse functions of the RCCs but also identifies different strategies for creating RCCs based on linguistic forms.

Release of experimental stimuli and questions for evaluating facial expressions in animations of American Sign Language

Matt Huenerfauth and Hernisa Kacorri

We have developed a collection of stimuli (with accompanying comprehension questions and subjective-evaluation questions) that can be used to evaluate the perception and understanding of facial expressions in ASL animations or videos. The stimuli have been designed as part of our laboratory's on-going research on synthesizing ASL facial expressions such as Topic, Negation, Yes/No Questions, WH-questions, and RH-questions. This paper announces the release of this resource, describes the collection and its creation, and provides sufficient details to enable researchers determine if it would benefit their work. Using this collection of stimuli and questions, we are seeking to evaluate computational models of ASL animations with linguistically meaningful facial expressions, which have accessibility applications for deaf users.

Computer-based tracking, analysis, and visualization of linguistically significant nonmanual events in American Sign Language (ASL)

Carol Neidle, Jingjing Liu, Bo Liu, Xi Peng, Christian Vogler and Dimitris Metaxas

Our linguistically annotated American Sign Language (ASL) corpora have formed a basis for research to automate detection by computer of essential linguistic information conveyed through facial expressions and head movements. We have tracked head position and facial deformations, and used computational learning to discern specific grammatical markings. Our ability to detect, identify, and temporally localize the occurrence of such markings in ASL videos has recently been improved by incorporation of (1) new techniques for deformable model-based 3D tracking of head position and facial expressions, which provide significantly better tracking accuracy and recover quickly from temporary loss of track due to occlusion; and (2) a computational learning approach

incorporating 2-level Conditional Random Fields (CRFs), suited to the multi-scale spatio-temporal characteristics of the data, which analyses not only low-level appearance characteristics, but also the patterns that enable identification of significant gestural components, such as periodic head movements and raised or lowered eyebrows. Here we summarize our linguistically motivated computational approach and the results for detection and recognition of nonmanual grammatical markings; demonstrate our data visualizations, and discuss the relevance for linguistic research; and describe work underway to enable such visualizations to be produced over large corpora and shared publicly on the Web.

Session B: Annotation

Saturday 31 May, 11:00 – 13:00
Chairperson: Jette Kristoffersen Poster Session

Mouth features as non-manual cues for the categorization of lexical and productive signs in French Sign Language (LSF)

Antonio Balvet and Marie-Anne Sallandre

In this paper, we present evidence from a case study in LSF, conducted on narratives from 6 adult signers. In this study, picture and video stimuli have been used in order to identify the role of non-manual features such as gaze, facial expressions and mouth features. Hereafter, we discuss the importance of mouth features as markers of the alternation between frozen (Lexical Units, LU) and productive signs (Highly Iconic Structures, HIS). Based on qualitative and quantitative analysis, we propose to consider mouth features, i.e. mouthings on the one hand, and mouth gestures on the other hand, as markers, respectively, of LU versus HIS. As such, we propose to consider mouthings and mouth gestures as fundamental cues for determining the nature, role and interpretation of manual signs, in conjunction with other non-manual features. We propose an ELAN annotation template for mouth features in Sign Languages, together with a discussion on the different mouth features and their respective roles as discourse and syntactic-semantic operators.

Eye gaze annotation practices: Description vs. interpretation

Annelies Braffort

If sharing best practices and conventions for annotation of Sign Language corpora is a growing activity, less attention has been given to the annotation of non-manual activity. This paper focuses on annotation of eye gaze. The aim is to report some of the practices, and begin a discussion on this topic, to be continued during the workshop. After having presented and discussed the nature of the annotation values in several projects, and having explained our own practices, we examine the level of interpretation in the annotation process, and how the design of annotation conventions can be motivated by limitations in the available annotation tools.

An annotation scheme for mouth actions in sign languages

Onno Crasborn and Richard Bank

This paper describes the annotation scheme that has been used for research on mouth actions in the Corpus NGT. An orthographic representation of the visible part of the mouthing is supplemented by the citation form of the word, a categorisation of the type of the mouth action, the number of syllables in the mouth action, (non)alignment of a corresponding sign, and a layer representing some special functions. The scheme has been used for a series of studies on Sign Language of the Netherlands. The structure and vocabularies for the annotation scheme are described, as well as the

experiences in its use so far. Annotations will be published in the second release of the Corpus NGT annotations in late 2014.

A hybrid formalism to parse sign languages

Rémi Dubot and Christophe Collet

Sign Language (SL) linguistic is dependent on the expensive task of annotating. Some automation is already available for low-level information (eg. body part tracking) and the lexical level has shown significant progresses. The syntactic level lacks annotated corpora as well as complete and consistent models. This article presents a solution for the automatic annotation of SL syntactic elements. It exposes a formalism able to represent both constituency-based and dependency-based models. The first enables the representation of structures one may want to annotate, the second aims at fulfilling the holes of the first. A parser is presented and used to conduct two experiments on the solution. One experiment is on a real corpus, the other is on a synthetic corpus.

Annotation of mouth activities with iLex

Thomas Hanke

In a purely bottom-up approach an annotation practice used for mouth activities would try to describe the phenomena and leave it to a second step to classify (e.g. between mouthing and mouth gestures) and relate (e.g. to spoken language words). For practical reasons, however, the first step is often skipped, and separate coding systems are applied to what is categorised either as mouthing derived from spoken language or mouth gesture where there is no obvious connection between the meaning expressed and any spoken language words expressing that same meaning. This happens not only for time (=budget) reasons, but also because it is difficult for coders to describe mouth visemes precisely if the sign/mouth combo already suggests what is to be seen on the mouth. While there are established coding procedures to avoid influence as far as possible (like only showing the signer's face, provided video quality is good enough), they make the approach very time-consuming, even if not counting quality assurance measures like inter-transcriber agreement. Some projects undertaken at the IDGS in Hamburg therefore leave it with a spoken-language-driven approach: The mouth activity is classified as either mouth gesture or mouthing, and in the latter case the German word is noted down that a competent DGS signer "reads" from the lips, i.e. that word from the set of words to be expected with the co-temporal sign in its context that matches the observation. Standard orthography is used unless there is a substantial deviation. For mouth gestures, holistic labels are used. These two extremes span a whole spectrum of coding approaches that can be used for mouth activities. The poster shows how iLex, the Hamburg sign language annotation workbench, supports the whole range of solutions from more time-series-like systems to those evaluating co-occurrence and semantic relatedness, from novice-friendly decision trees to expert-only modes.

Mouth-based non-manual coding schema used in the Auslan corpus: Explanation, application and preliminary results

Trevor Johnston and Jane van Roekel

We describe a corpus-based study of one type of non-manual in signed languages (SLs) — mouth actions. Our ultimate aim is to examine the distribution and characteristics of mouth actions in Auslan (Australian Sign Language) to gauge the degree of language-specific conventionalization of these forms. We divide mouth gestures into categories broadly based on Crasborn et al. (2008), but modified to accommodate our experiences with the Auslan data. All signs and all mouth actions are examined and the state of the mouth in each sign is assigned to one of three broad categories: (i) mouthings, (ii) mouth gestures, and (iii) no mouth action. Mouth actions that invariably occur while

communicating in SLs have posed a number of questions for linguists: which are 'merely borrowings' from the relevant ambient spoken language (SpL)? Which are gestural and shared with all of the members of the wider community in which signers find themselves? And which are conventionalized aspects of the grammar of some or all SLs? We believe this schema captures all the relevant information about mouth forms and their use and meaning in context to enable us to describe their function and degree of conventionality.

Signing thoughts! A methodological approach within the semantic field work used for coding nonmanuals which express modality in Austrian Sign Language (ÖGS)

Andrea Lackner and Nikolaus Riemer

Signing thoughts gives the possibility to express unreal situations, possibilities and so forth. Also, signers may express their attitude on these thoughts such as being uncertain about an imagined situation. We describe a methodological approach within the semantic field work which was used for identifying nonmanuals in Austrian Sign Language (ÖGS) which tend to occur in thoughts and which may code (epistemic and deontic) modality. First, the process of recording short stories which most likely include lines of thoughts is shown. Second, the annotation process and the outcome of this process is described. The findings show that in almost all cases the different annotators identified the same non-manual movements/positions and the same starting and ending points of these nonmanuals in association with the lexical entries. The direction of motion was distinguished by a contrast of movement. Some nonmanuals were distinguished due to intensified performance, size of performance, speed of performance, an additional movement component, or additional body tension. Finally, we present nonmanuals which frequently occur in signed thoughts. These include various epistemic markers, a deontic marker, indicators which show the hypothetical nature of signed thoughts, and an interrogative marker which is different to interrogative markers in direct questions.

Addressing the cardinals puzzle: New insights from non-manual markers in Italian Sign Language

Lara Mantovan, Carlo Geraci and Anna Cardinaletti

This paper aims at investigating the main linguistic properties associated with cardinal numerals in LIS (Italian Sign Language). Considering this issue from several perspectives (phonology, prosody, semantics and syntax), we discuss some relevant corpus and elicited data with the purpose of shedding light on the distribution of cardinals in LIS. We also explain what triggers the emergence of different word/sign orders in the noun phrase. Non-manual markers are crucial in detecting the two particular subcases.

Taking non-manuality into account in collecting and analyzing Finnish Sign Language video data

Anna Puupponen, Tommi Jantunen, Ritva Takkinen, Tuija Wainio and Outi Pippuri

This paper describes our attention to research into non-manuals when collecting a large body of video data in Finnish Sign Language (FinSL). We will first of all give an overview of the data-collecting process and of the choices that we made in order for the data to be usable in research into non-manual activity (e.g. camera arrangement, video compression, and Kinect technology). Secondly, the paper will outline our plans for the analysis of the non-manual features of this data. We discuss the technological methods we plan to use in our investigation of non-manual features (i.e. computer-vision based methods) and give examples of the type of results that this kind of approach can provide us with.

Visualizing the spatial working memory in mathematical discourse in Finnish Sign Language

Päivi Rainò, Marja Huovila and Irja Seilola

In this paper, we will present problems that arise when trying to render legible signed texts containing mathematical discourse in Finnish Sign Language. Calculation processes in sign language are carried out using fingers, both hands and the three-dimensional neutral space in front of the signer. Specific hand movements and especially the space in front of the body function like a working memory where fingers, hands and space are used as buoys in a regular and syntactically well-defined manner when retrieving, for example, subtotals. As these calculation processes are performed in fragments of seconds with both hands that act individually, simultaneity and multidimensionality create problems for traditional coding and notation systems used in sign language research. Conversion to glosses or translations to spoken or written text (e.g. in Finnish or English) has proven challenging and what is most important, none of these ways gives justice to this unique concept mapping and mathematical thinking in signed language. Our proposal is an intermediary solution, a simple numeric animation while looking for a more developed, possibly a three-dimensional representation to visualise the calculation processes in signed languages.

Use of nonmanuals by adult L2 signers in Swedish Sign Language – Annotating the nonmanuals

Krister Schönström and Johanna Mesch

Nonmanuals serve as important grammatical markers for different syntactic constructions, e.g. marking clause types. To account for the acquisition of syntax by L2 SSL learners, therefore, we need to have the ability to annotate and analyze nonmanual signals. Despite their significance, however, these signals have yet to be the topic of research in the area of SSL as an L2. In this paper, we will provide suggestions for annotating the nonmanuals in L2 SSL learners. Data is based on a new SSL as L2 corpus from our on-going project entitled "L2 Corpus in Swedish Sign Language." In this paper, the combination of our work in grammatical analysis and in the creation of annotating standards for L2 nonmanuals, as well as preliminary results from the project, will be presented.

Session C: Language Technology

Saturday 31 May, 14:00 – 16:00
Chairperson: Johanna Mesch Poster Session

Synthesizing facial expressions for sign language avatars

Yosra Bouzid, Oussama El Ghoul and Mohamed Jemni

Sign language is more than just moving the fingers or hands; it is a visual language in which non manual gestures play a very important role. Recently, a growing body of research has paid increasing attention to the development of signing avatars endowed with a set of facial expressions in order to perform the actual functioning of the sign language, and gain wider acceptance by deaf users. In this paper, we propose an effective method to generate facial expressions for signing avatars based on the physics-based muscle model. The main focus of our work is to automate the task of the muscle mapping on the face model in the correct anatomical positions and the detection of the jaw part by using a small set of MPEG-4 Feature Points of the given mesh.

Implementation of an automatic sign language lexical annotation framework based on Propositional Dynamic Logic

Arturo Curiel and Christophe Collet

In this paper, we present the implementation of an automatic Sign Language (SL) sign annotation framework based on a formal logic, the Propositional Dynamic Logic (PDL). Our system relies heavily on the use of a specific variant of PDL, the Propositional Dynamic Logic for Sign Language (PDL_{SL}), which lets us describe SL signs as formulae and corpora videos as labelled transition systems (LTSs). Here, we intend to show how a generic annotation system can be constructed upon these underlying theoretical principles, regardless of the tracking technologies available or the input format of corpora. With this in mind, we generated a development framework that adapts the system to specific use cases. Furthermore, we present some results obtained by our application when adapted to one distinct case, 2D corpora analysis with pre-processed tracking information. We also present some insights on how such a technology can be used to analyze 3D real-time data, captured with a depth device.

Creation of a multipurpose sign language lexical resource: The GSL lexicon database

Athanasia-Lida Dimou, Theodore Goulas, Eleni Efthimiou, Stavroula-Evita Fotinea, Panagiotis Karioris, Michalis Pissaris, Dimitris Korakakis and Kiki Vasilaki

The GSL lexicon database is the first extensive database of Greek Sign Language (GSL) signs, created on the basis of knowledge derived from the linguistic analysis of natural signers' data. It incorporates a lemma list that currently includes approximately 6,000 entries and is intended to reach a total number of 10,000 entries within the next two years. The design of the database allows for classification of signs on the basis of their articulation features as regards both manual and non-manual elements. The adopted information management schema accompanying each entry provides for retrieval according to a variety of linguistic properties. In parallel, annotation of the full set of sign articulation features feeds more natural performance of synthetic signing engines and more effective treatment of sign language (SL) data in the framework of sign recognition and natural language processing.

When non-manuals meet semantics and syntax: Towards a practical guide for the segmentation of sign language discourse

Sílvia Gabarró-López and Laurence Meurant

This paper aims to contribute to the segmentation of sign language (SL) discourses by providing an operational synthesis of the criteria that signers use to segment a SL discourse. Such procedure was required when it came to analyse the role of buoys as discourse markers (DMs), which is part of a PhD on DMs in French Belgian SL (LSFB). All buoy markers found in the data had to be differentiated in terms of scope: some markers (like most list buoy markers) seemed to be long-range markers, whereas others (like most fragment buoy markers) seemed to have a local scope only. Our practical guide results from a hierarchized and operationalized synthesis of the criteria, which explain the segmentation judgments of deaf (native and non-native) and hearing (non-native) signers of LSFB who were asked to segment a small-scale (1h) corpus. These criteria are a combination of non-manual, semantic and syntactic cues. Our contribution aims to be shared, tested on other SLs and hopefully improved to provide SL researchers who conduct discourse studies with some efficient and easy-to-use guidelines, and avoid them extensive (and time-consuming) annotation of the manual and non-manual cues that are related to the marking of boundaries in SLs.

Last train to "Rebaudengo Fossano": The case of some names in avatar translation

Carlo Geraci and Alessandro Mazzei

In this study, we present an unorthodox case study where cross-linguistic and cross modal information is provided by a "non-manual" channel during the process of automatic translation from spoken into sign language (SL) via virtual actors (avatars). Specifically, we blended written forms (crucially, not subtitles) into the sign stream in order to import the names of less-known train stations into Italian Sign Language (LIS). This written Italian-LIS blending is a more effective compromise for Deaf passengers than fully native solutions like fingerspelling or using the local less-known SL names. We report here on part of an on-going project, LIS4ALL, aiming at producing a prototype avatar signing train station announcements. The final product will be exhibited at the train station of Torino Porta Nuova in Turin, Italy.

How to use depth sensors in sign language corpus recordings

Rekha Jayaprakash and Thomas Hanke

Recently, combined camera and depth sensor devices caused substantial advances in Computer Vision directly applicable to automatic coding a signer's use of head movement, eye gaze, and, to some extent, facial expression. Automatic and even semi-automatic annotation of nonmanuals would mean dramatic savings on annotation time and are therefore of high interest for anyone working on sign language corpora. Optimally, these devices need to be placed directly in front of the signer's face at a distance of less than 1m. While this might be ok for some experimental setups, it is definitely nothing to be used in a corpus setting for at least two reasons: (i) The signer looks at the device instead of into the eyes of an interlocutor. (ii) The device is in the field of view of other cameras used to record the signer's manual and nonmanual behaviour. Here we report on experiments determining the degradation in performance when moving the devices away from their optimal positions in order to achieve a recording setup acceptable in a corpus context. For these experiments, we used two different device types (Kinect and Carmine 1.09) in combination with one mature CV software package specialised on face recognition (FaceShift). We speculate about the reasons for the asymmetries detected and how they could be resolved. We then apply the results to the studio setting used in the DGS Corpus project and show how the signers' and cameras fields of view are influenced by introducing the new devices and we are happy to discuss the acceptability of this approach.

Weakly supervised automatic transcription of mouthings for gloss-based sign language corpora

Oscar Koller, Hermann Ney and Richard Bowden

In this work we propose a method to automatically annotate mouthings in sign language corpora, requiring no more than a simple gloss annotation and a source of weak supervision, such as automatic speech transcripts. For a long time, research on automatic recognition of sign language has focused on the manual components. However, a full understanding of sign language is not possible without exploring its remaining parameters. Mouthings provide important information to disambiguate homophones with respect to the manuals. Nevertheless most corpora for pattern recognition purposes are lacking any mouthing annotations. To our knowledge no previous work exists that automatically annotates mouthings in the context of sign language. Our method produces a frame error rate of 39% for a single signer.

Estimating head pose and state of facial elements for sign language video

Marcos Luzardo, Ville Viitaniemi, Matti Karppa, Jorma Laaksonen and Tommi Jantunen

In this work we present methods for automatic estimation of non-manual gestures in sign language videos. More specifically, we study the estimation of three head pose angles (yaw, pitch, roll) and the state of facial elements (eyebrow position, eye openness, and mouth state). This kind of estimation facilitates automatic annotation of sign language videos and promotes more prolific production of annotated sign language corpora. The proposed estimation methods are incorporated in our publicly available SLMotion software package for sign language video processing and analysis. Our method implements a model-based approach: for head pose we employ facial landmarks and skin masks as features, and estimate yaw and pitch angles by regression and roll using a geometric measure; for the state of facial elements we use the geometric information of facial elements of the face as features, and estimate quantized states using a classification algorithm. We evaluate the results of our proposed methods in quantitative and qualitative experiments.

The "how-to" of integrating FACS and ELAN for analysis of non-manual features in ASL

Kristin Mulrooney, Julie Hochgesang, Carla Morris and Katie Lee

The process of transcribing and annotating non-manual features presents challenges for sign language researchers. This paper describes the approach used by our research team to integrate the Facial Action Coding System (FACS) with the EUDICO Linguistic Annotator (ELAN) program to allow us to more accurately and efficiently code non-manual features. Preliminary findings are presented which demonstrate that this approach is useful for a fuller description of facial expressions.

Non-manuals and markers of (dis)fluency

Ingrid Notarrigo and Laurence Meurant

This paper focuses on the analysis and annotation of non-manual features in the framework of a study of (dis)fluency markers in French Belgian Sign Language (LSFB). In line with Götz (2013), we consider (dis)fluency as the result of the combination of many independent markers ('fluencemes'). These fluencemes may contribute either positively or negatively to the efficiency of a discourse depending on their context of appearance, their specific combination, their position and frequency. We show that the non-manual features in LSFB make distinctions within pauses and palm-up signs in a consistent way and contribute to the value of the manual marker. The selection of a limited number of relevant combinations of nonmanuals, in the context of pauses and palm-up signs, proves to simplify the annotation process and to limit the number of features to examine for each nonmanual. The gaze and the head appear to be necessary and sufficient to describe pauses and palm-up signs accurately. Though these findings are limited to this pilot study, they will pave the way to the next steps of the broader research project on (dis)fluency markers in LSFB this work is part of.

Session D: Integration

Saturday 31 May, 16:30 – 18:00
Chairperson: Onno Crasborn Oral/Signed Session

Analysis for synthesis: Investigating corpora for supporting the automatic generation of role shift

John McDonald, Rosalee Wolfe, Robyn Moncrief and Souad Baowidan

In signed languages, role shift is a process that can facilitate the description of statements, actions or thoughts of someone other than the person who is signing, and sign synthesis systems must be able to automatically create animations that portray it effectively. Animation is only as good as the data used to create it, which is the motivation for using corpus analyses when developing new tools and techniques. This paper describes work-in-progress towards automatically generating role shift in discourse. This effort includes consideration of the underlying representation necessary to generate a role shift automatically and a survey of current annotation approaches to ascertain whether they supply sufficient data for the representation to generate the role shift.

Non-manual features: The right to indifference

Michael Filhol, Mohamed Nassime Hadjadj and Annick Choisier

This paper discusses the way sign language can be described with a global account of the visual channel, not separating manual articulators in any way. In a first section, it shows that non-manuals are often either ignored in favour of manual focus, or included but given roles that are mostly different from the mainly hand-assigned lexical role. A second section describes the AZee model as a tool to describe Sign Language productions without assuming any separation, neither between articulators nor between grammatical roles. We conclude by giving a full AZee description for one of the several examples populating the paper.

Segmenting the Swedish Sign Language corpus: On the possibilities of using visual cues as a basis for syntactic segmentation

Carl Börstell, Johanna Mesch and Lars Wallin

This paper deals with the possibility of conducting syntactic segmentation of the Swedish Sign Language Corpus (SSLC) on the basis of the visual cues from both manual and nonmanual signals. The SSLC currently features segmentation on the lexical level only, which is why the need for a linguistically valid segmentation on e.g. the clausal level would be very useful for corpus-based studies on the grammatical structure of Swedish Sign Language (SSL). An experiment was carried out letting seven Deaf signers of SSL each segment two short texts (one narrative and one dialogue) using ELAN, based on the visual cues they perceived as boundaries. This was later compared to the linguistic analysis done by a language expert (also a Deaf signer of SSL), who segmented the same texts into what was considered syntactic clausal units. Furthermore, these segmentation procedures were compared to the segmentation done for the Swedish translations also found in the SSLC. The results show that though the visual and syntactic segmentations overlap in many cases, especially when a number of cues coincide, the visual segmentation is not consistent enough to be used as a means of segmenting syntactic units in the SSLC.

Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing

31 May 2014

ABSTRACTS

Editors:
Sophia Ananiadou, Khalid Choukri, Kevin Bretonnel Cohen, Dina Demner-Fushman, Jan Hajic, Allan Hanbury, Gareth Jones, Henning Müller, Pavel Pecina and Paul Thompson

Workshop Programme

09:15 – 09:30 Welcome

09:30 – 10:30 Invited presentation: Georgios Paliouras, *The BioASQ Project*

10:30 – 11:00 Coffee break

11:00 – 11:20 Rachel Ginn, Pranoti Pimpalkhute, Azadeh Nikfarjam, Apurv Patki, Karen O'Connor, Abeed Sarker, Karen Smith and Graciela Gonzalez, *Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark*

11:20 – 11:40 Takashi Okumura, Eiji Aramaki and Yuka Tateisi, *Expression of Laboratory Examination Results in Medical Literature*

11:40 – 12:00 Erwin Marsi, Pinar Öztürk, Elias Aamot, Gleb Sizov and Murat V. Ardelan, *Towards Text Mining in Climate Science: Extraction of Quantitative Variables and their Relations*

12:00 – 12:20 Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset and Pierre Zweigenbaum, *The Quaero French Medical Corpus: A Resource for Medical Entity Recognition and Normalization*

12:20 – 14:00 Lunch break

14:00 – 14:20 Rezarta Islamaj Doğan, W. John Wilbur and Donald C. Comeau, *BioC and Simplified Use of the PMC Open Access Dataset for Biomedical Text Mining*

14:20 – 14:40 Kirk Roberts, Kate Masterton, Marcelo Fiszman, Dina Demner-Fushman and Halil Kilicoglu, *Annotating Question Types for Consumer Health Questions*

14:40 – 15:10 Short Poster Presentations

Xiao Fu and Sophia Ananiadou, *Improving the Extraction of Clinical Concepts from Clinical Records*

Kerstin Denecke, *Extracting Medical Concepts from Medical Social Media with Clinical NLP tools: A Qualitative Study*

Mariana Neves, Konrad Herbst, Matthias Uflacker and Hasso Plattner, *Preliminary Evaluation of Passage Retrieval in Biomedical Multilingual Question Answering*

Borbála Siklósi, Attila Novák and Gábor Prószéky, *Resolving Abbreviations in Clinical Texts without Pre-Existing Structured Resources*

Lina Henriksen, Anders Johannsen, Bart Jongejan, Bente Maegaard and Jürgen Wedekind, *Worlds Apart - Ontological Knowledge in Question Answering for Patients*

Behrouz Bokharaeian, Alberto Díaz, Mariana Neves and Virginia Francisco, *Exploring Negation Annotations in the DrugDDI Corpus*

Stefan Schulz, Catalina Martinez Costa, Markus Kreuzthaler, Jose Antonio Miñarro-Giménez, Ulrich Andersen, Anders Boeck Jensen and Bente Maegaard, *Semantic Relation Discovery by using Co-occurrence Information*

Lorraine Goeuriot, Wendy Chapman, Gareth J.F. Jones, Liadh Kelly, Johannes Leveling and Sanna Salanterä, *Building Realistic Potential Patients Queries for Medical Information Retrieval Evaluation*

15:10 – 16:30 Poster Session (Continuing into the Coffee Break)

16:00 – 16:30 Coffee break

16:30 – 18:00 Structured Discussion

- What are the most needed resources for health and biomedical text processing that currently don't exist or are not available?
- What would their availability enable us to do that we can't do now?
- Why are these resources not available or don't exist?
- How can we make them available or create them?

Workshop Organisers

Sophia Ananiadou	University of Manchester, UK
Khalid Choukri	ELDA, France
Kevin Bretonnel Cohen	University of Colorado, USA
Dina Demner-Fushman	National Library of Medicine, USA
Jan Hajic	Charles University Prague, Czech Republic
Allan Hanbury	Technical University of Vienna, Austria
Gareth Jones	Dublin City University, Ireland
Henning Müller	HES-SO Valais, Switzerland
Pavel Pecina	Charles University Prague, Czech Republic
Paul Thompson	University of Manchester, UK

Workshop Programme Committee

Olivier Bodenreider	National Library of Medicine, USA
Wendy Chapman,	University of Utah, USA
Hercules Dalianis	University of Stockholm, Sweden
Noémie Elhadad	Columbia University, USA
Graziela Gonzalez	Arizona University, USA
Jin-Dong Kim	DBCLS, Japan
Dimitris Kokkinakis	Gothenburg University, Sweden
Ioannis Korkontzelos	University of Manchester, UK
Hongfang Liu	Mayo Clinic, USA
Naoaki Okazaki	Tohoku University, Japan
Arzucan Özgür	Bogazici University, Turkey
Claire Nedellec	INRA, France
Sampo Pyysalo	University of Turku, Finland
Fabio Rinaldi	University of Zurich, Switzerland
Andrey Rzhetsky	University of Chicago, USA
Guergana Savova	Children's Hospital Boston and Harvard Medical School, USA
Hagit Shatkay	University of Delaware, USA
Rafal Rak	University of Manchester, UK
Lucy Vanderwende	Microsoft, USA
Karin Verspoor	NICTA, Australia
John Wilbur	NCBI, NLM, NIH, USA
Stephen Wu	Mayo Clinic, USA
Pierre Zweigenbaum	LIMSI, France

Introduction

This volume contains the papers accepted at the *4rd Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing (BioTxtM 2014)*, held at LREC 2014, Reykjavik. Over the past years, there has been an exponential growth in the amount of biomedical and health information available in digital form. In addition to the 23 million references to biomedical literature currently available in PubMed, other sources of information are becoming more readily available. For example, there is a wealth of information available in clinical records, whilst the growing popularity of social media channels has resulted in the creation of various specialised groups. Extensive information is also available in languages other than English.

With such a deluge of information at their fingertips, domain experts and health professionals have an ever-increasing need for tools that can help them to isolate relevant nuggets of information in a timely and efficient manner, regardless of both information source and mother tongue. However, this goal presents many new challenges in analysis and search. For example, given the highly multilingual nature of available information, it is important that language barriers do not result in vital information being missed. In addition, different information sources cover varying topics and contain differing styles of language, while varying terminology may be used by lay persons, academics and health professionals. There is also often little standardisation amongst the extensive use of abbreviations found in medical and health-related text.

Applications in the health and biomedical domain are reliant on high quality resources. These include databases and ontologies (e.g., Biothesaurus, UMLS Metathesaurus) and lexica (e.g., BioLexicon and UMLS SPECIALIST lexicon). Given the frequently changing and variable nature of biomedical terminology and abbreviations, combined with the requirement to take multilingual information into account, there is an urgent need to investigate new ways of creating, updating such resources, or adapting them to new languages. New techniques may include combining semi-automatic methods, machine translation techniques, crowdsourcing or other collaborative efforts.

Community shared tasks and challenges (e.g. Biocreative I-IV, ACL BioNLP Shared Tasks (2009-2011-2013) etc.) have resulted in an increase in the number of annotated corpora, covering an ever-expanding range of sub-domains and annotation types. Such corpora are helping to steer research efforts to focus on open research problems, as well as encouraging the development of increasingly adaptable and wider coverage text mining tools. Interoperability and reuse are also vital considerations, as evidenced by efforts such as the BioCreative Interoperability Initiative (BioC) and the UIMA architecture. Several of the corpora introduced above are compliant with both BioC and UIMA, and are available within the U-Compare and Argo systems, which allow easy construction of NLP workflows and evaluation against gold standard corpora. There is also a need to consider how resources and techniques can facilitate easier access to relevant information that is written in a variety of different languages. For example, can existing techniques and resources used for machine translation, multilingual search and question answering in other domains be adapted simplify access to multilingual information in the biomedical and health domains?

The papers in this volume exemplify the diversity of research that is currently taking place, and explain how some of the challenges introduced above are being addressed. A number of research topics involving clinical corpora are represented, including the extraction of concepts (Fu and Ananiadou), resolution of abbreviations (Siklós et al.) and automatic generation of queries from medical reports as a means of enhancing patients' understanding of eHealth data (Goeuriot et al). Research into helping patients and consumers to obtain answers to medical queries is also the topic of a number of other papers, including the classification of question types as a means to determine the best strategy for answer selection (Roberts et al.), and the exploration of co-occurrence data in

UMLS to infer non-ontological semantic relations for the construction a knowledge base to support patient-centred question answering (Schulz et al.). Henriksen et al. describe a system operating on the Danish language that answers patients' queries by combining the use of formalised knowledge from a different medical resource (SNOMED CT) with text-to-text generation in the form of document summarisation and question generation. Also in the area of question answering, Neves et al. present a multilingual system for biomedical literature, operating on English, German and Spanish. A new corpus of biomedical documents in French annotated at the entity and concept level (Névéol et al.) further demonstrates the increasing efforts to facilitate the development of domain-specific systems in languages other than English. Okumur et al. explore the literature from a different perspective, i.e., as a source of information about laboratory results for use in building a disease knowledge base. Completing the set of papers concerning biomedical literature, Doğan et al. addresses the important issue of interoperability, through the description of the BioC annotation format, and the introduction of the BioC-PMC dataset, which contains annotations that conform to this description.

Exploiting the valuable knowledge available within the increasing volume of social media data is the subject of two papers. One explores the extraction of medical concepts from medical social media (Denecke), whilst the other describes a new corpus of tweets annotated with adverse drug reactions (Ginn et al.). The importance of studying the effects of drugs is also highlighted by Bokharaeian et al., who have enriched the DDI corpus, containing information about interactions between drugs, with negation cues and scopes. Finally, taking inspiration from research into biomedical literature, Marsi et al. explore the first steps in extracting entities and events from neighbouring research fields, i.e., climate science, marine science and environmental science, through the design of a new annotation scheme.

We wish to thank the authors for submitting papers for consideration, and the members of the programme committee for offering their time and effort to review the submissions. We would also like to thank our invited speaker, Dr. Georgios Paliouras, for his contribution.

Sophia Ananiadou, Khalid Choukri, Kevin Bretonnel Cohen, Dina Demner-Fushman, Jan Hajic, Allan Hanbury, Gareth Jones, Henning Müller, Pavel Pecina and Paul Thompson

Mining Twitter for Adverse Drug Reaction Mentions: A Corpus and Classification Benchmark

Rachel Ginn, Pranoti Pimpalkhute, Azadeh Nikfarjam, Apurv Patki, Karen O'Connor, Abeed Sarker, Karen Smith and Graciela Gonzalez

With many adults using social media to discuss health information, researchers have begun diving into this resource to monitor or detect health conditions on a population level. Twitter, specifically, has flourished to several hundred million users and could present a rich information source for the detection of serious medical conditions, like adverse drug reactions (ADRs). However, Twitter also presents unique challenges due to brevity, lack of structure, and informal language. We present a freely available, manually annotated corpus of 10,822 tweets, which can be used to train automated tools to mine Twitter for ADRs. We collected tweets utilizing drug names as keywords, but expanding them by applying an algorithm to generate misspelled versions of the drug names for maximum coverage. We annotated each tweet for the presence of a mention of an ADR, and for those that had one, annotated the mention (including span and UMLS IDs of the ADRs). Our inter-annotator agreement for the binary classification had a Kappa value of 0.69, which may be considered substantial (Viera & Garrett, 2005). We evaluated the utility of the corpus by training two classes of machine learning algorithms: Naïve Bayes and Support Vector Machines. The results we present validate the usefulness of the corpus for automated mining tasks. The classification corpus is available from http://diego.asu.edu/downloads.

Expression of Laboratory Examination Results in Medical Literature

Takashi Okumura, Eiji Aramaki and Yuka Tateisi

Medical literature contains expressions of laboratory examination results, which are invaluable knowledge sources for building a disease knowledge base that covers even rare diseases. In this study, we analyzed such expressions of disease descriptions in open databases with manually built dictionaries and obtained the following results. First, we identified two major types of expressions for laboratory examination results, that with and without their test names in the expressions. Second, the study identified evaluative expressions that frequently appear in the description of the results. Third, presence of test names and evaluative expressions could classify the expressions into four major classes that demand independent strategies to interpret. The study illustrated that this is beyond the scope of the existing corpora in this domain mostly designed for medical records. Although the analysis is based on rudimentary statistics, it clarified the factors necessary for future corpus designs to promote further research.

Towards Text Mining in Climate Science: Extraction of Quantitative Variables and their Relations

Erwin Marsi, Pinar Öztürk, Elias Aamot, Gleb Sizov and Murat V. Ardelan

This paper addresses text mining in the cross-disciplinary fields of climate science, marine science and environmental science. It is motivated by the desire for literature-based knowledge discovery from scientific publications. The particular goal is to automatically extract relations between quantitative variables from raw text. This results in rules of the form "If variable X increases, than variable Y decreases". As a first step in this direction, an annotation scheme is proposed to capture

the events of interest – those of change, cause, correlation and feedback – and the entities involved in them, quantitative variables. Its purpose is to serve as an intermediary step in the process of rule extraction. It is shown that the desired rules can indeed be automatically extracted from annotated text. A number of open challenges are discussed, including automatic annotation, normalisation of variables, reasoning with rules in combination with domain knowledge and the need for meta-knowledge regarding context of use.

The Quaero French Medical Corpus: A Resource for Medical Entity Recognition and Normalization

Aurélie Névéol, Cyril Grouin, Jeremy Leixa, Sophie Rosset and Pierre Zweigenbaum

A vast amount of information in the biomedical domain is available as natural language free text. An increasing number of documents in the field are written in languages other than English. Therefore, it is essential to develop resources, methods and tools that address Natural Language Processing in the variety of languages used by the biomedical community. In this paper, we report on the development of an extensive corpus of biomedical documents in French annotated at the entity and concept level. Three text genres are covered, comprising a total of 103,056 words. Ten entity categories corresponding to UMLS Semantic Groups were annotated, using automatic pre-annotations validated by trained human annotators. The pre-annotation method was found helpful for entities and achieved above 0.83 precision for all text genres. Overall, a total of 26,409 entity annotations were mapped to 5,797 unique UMLS concepts.

BioC and Simplified Use of the PMC Open Access Dataset for Biomedical Text Mining

Rezarta Islamaj Doğan, W. John Wilbur and Donald C. Comeau

High quality easily-accessible resources are crucial for developing reliable applications in the health and biomedical domain. At the same time, interoperability, broad use, and reuse are vital considerations when developing useful systems. As a response, BioC has recently been put forward as a convenient XML format to share text documents and annotations, and as an accompanying input/output library to promote interoperability of data and tools. The BioC approach allows a large number of different textual annotations to be represented, and permits developers to more easily and efficiently share training data, supportive software modules and produced results. Here we give a brief overview of BioC resources. We also present the BioC-PMC dataset as a new resource, which contains all the articles available from the PubMed Central Open Access collection conveniently packaged in the BioC format. We show how this valuable resource can be easily used for text-mining tasks. Code and data are available for download at the BioC site: http://bioc.sourceforge.net.

Annotating Question Types for Consumer Health Questions

Kirk Roberts, Kate Masterton, Marcelo Fiszman, Dina Demner-Fushman and Halil Kilicoglu

This paper presents a question classification scheme and a corresponding annotated corpus of consumer health questions. While most medical question classification methods have targeted medical professionals, the 13 question types we present are targeted toward disease questions posed by consumers. The corpus consists of 1,467 consumer-generated requests for disease information,

containing a total of 2,937 questions. The goal of question type classification is to indicate the best strategy for selecting answers from publicly available health information resources, such as medical encyclopedias and medical research websites. The annotated question corpus, along with detailed annotation guidelines, are publicly available.

Improving the Extraction of Clinical Concepts from Clinical Records

Xiao Fu and Sophia Ananiadou

Essential information relevant to medical problems, tests, and treatments is often expressed in patient clinical records with natural language, making their processing a daunting task for automated systems. One of the steps towards alleviating this problem is concept extraction. In this work, we proposed a machine learning-based named entity recognition system to extract clinical concepts from patient discharge summaries and progress notes without the need for any external knowledge resources. Three pre- and post-processing methods were investigated, i.e. truecasing, abbreviation disambiguation, and distributional thesaurus lookup, the individual annotation results of which were combined into a final annotation set using two refinement schemes. While truecasing and abbreviation disambiguation capture the inflectional morphology of words, the distributional thesaurus lookup allows for statistics-based similarity matching. We achieved a maximum F-score of 0.7586 and 0.8444 for exact and inexact matching, respectively. Our results show that truecasing and annotation combination are the enhancements which best increase the system performance, whereas abbreviation disambiguation and distributional thesaurus lookup bring about insignificant improvements.

Extracting Medical Concepts from Medical Social Media with Clinical NLP tools: A Qualitative Study

Kerstin Denecke

Medical social-media provides a rich source of information on diagnoses, treatment and experiences. For its automatic analysis, tools need to be available that are able to process this particular data. Since content and language of medical social-media differs from those of general social media and of clinical document, additional methods are necessary in particular to identify medical concepts and relations among them. In this paper, we analyse the quality of two existing tools for extracting clinical terms from natural language that were originally developed for processing clinical documents (cTakes, MetaMap) by applying them on a real-world set of medical blog postings. The results show that medical concepts that are explicitly mentioned in texts can reliably be extracted by those tools also from medical social-media data, but the extraction misses relevant information captured in paraphrase or formulated in common language.

Preliminary Evaluation of Passage Retrieval in Biomedical Multilingual Question Answering

Mariana Neves, Konrad Herbst, Matthias Uflacker and Hasso Plattner

Question answering systems can support biologists and physicians when searching for answers in the scientific literature or the Web. Further, multilingual question answering systems provide more possibilities and flexibility to users, by allowing them to write questions and get answers in their native language, and the exploration of resources in other languages by means of machine

translation. We present a prototype of a multilingual question answering system for the biomedical domain for English, German and Spanish. Preliminary experiments have been carried out for passage retrieval based on the multilingual parallel datasets of Medline titles released in the CLEF-ER challenge. Two parallel collections of 50 questions for English/German and English/Spanish have been created to support evaluation of the system. Results show that the task is not straightforward, that additional semantic resources are needed to support query expansion and that different strategies might be necessary for distinct languages.

Resolving Abbreviations in Clinical Texts without Pre-Existing Structured Resources

Borbála Siklósi, Attila Novák and Gábor Prószéky

One of the most important topics in clinical text processing is the identification of relevant concepts. This includes the detection and resolution of abbreviations, acronyms, or other shortened forms in the documents. Even though the task of resolving abbreviations can be treated as a word sense disambiguation problem, such methods require structured lexical knowledge bases. However, for less-resourced languages such resources are not available. In this paper, a method is proposed for the disambiguation and resolution of abbreviations found in Hungarian clinical records. In order to achieve reasonable performance, a lexicon must be created for each domain. It is shown how the set of entries to be included in such a lexicon can be induced from the corpus, thus the manual effort of creating a lexicon can be reduced significantly. The results for resolving abbreviations in Hungarian clinical documents are also shown, which are achieved by using the corpus instead of non-existing structured resources.

Worlds Apart - Ontological Knowledge in Question Answering for Patients

Lina Henriksen, Anders Johannsen, Bart Jongejan, Bente Maegaard and Jürgen Wedekind

We present ESICT, a hybrid question-answering system building on formalized knowledge from a medical ontology (SNOMED CT) as well as text-to-text generation in the form of document summarization and question generation. The independent subsystems are queried in parallel and compete for delivering the best answer. The use of an ontology gives the patient access to information typically not found in other sources, but also exposes a gap between everyday language and the specialized terms and conceptualizations of health professionals. In this paper we describe the ESICT system and discuss for each strategy how well it deals with this gap.

Exploring Negation Annotations in the DrugDDI Corpus

Behrouz Bokharaeian, Alberto Díaz, Mariana Neves and Virginia Francisco

Detecting drug-drug interactions (DDI) is an important research field in pharmacology and medicine and several publications report every year the negative effect of combining drugs and chemical treatments. The DrugDDI corpus is a collection of documents derived from the DrugBank database and contains manual annotations for interactions between drugs. We have investigated the negated statements in this corpus and found that they consist of approximately 21% of its sentences. Previous works have shown that considering features related to negation can improve results for the DDI task. The main goal of this paper is to describe the process for annotating the DDI-DrugBank corpus with negation cues and scopes, to show the correlations between these and the DDI

annotations and to demonstrate that negations can be used as features for a DDI detection system. Basic experiments have been carried out to show the benefits when considering negations in the DDI task. We believe that the extended corpus can be a significant progress in training and testing algorithms for DDI extraction.

Semantic Relation Discovery by using Co-occurrence Information

Stefan Schulz, Catalina Martinez Costa, Markus Kreuzthaler, Jose Antonio Miñarro-Giménez, Ulrich Andersen, Anders Boeck Jensen and Bente Maegaard

Motivated by the need of constructing a knowledge base for a patient-centred question-answering system, the potential of exploiting co-occurrence data to infer non-ontological semantic relations out of these statistical associations is explored. The UMLS concept co-occurrence table MRCOC is used as a data source. This data provides, for each co-occurrence record, a profile of MeSH subheading profiles. This is used as an additional source of semantic information from which we generate hypotheses for more specific semantic relations. An initial experiment was performed, limited to the study of disease-substance associations. For validation 20 diseases were selected and annotated by experts regarding treatment and prevention. The results showed good precision values (82 for prevention, 72 for treatment), but unsatisfactory values for recall (67 for prevention, 45 for treatment) for this particular use case.

Building Realistic Potential Patients Queries for Medical Information Retrieval Evaluation

Lorraine Goeuriot, Wendy Chapman, Gareth J.F. Jones, Liadh Kelly, Johannes Leveling and Sanna Salanterä

To evaluate and improve medical information retrieval, benchmarking data sets need to be created. Few benchmarks have been focusing on patients' information needs. There is a need for additional benchmarks to enable research into effective retrieval methods. In this paper we describe the manual creation of patient queries and investigate their automatic generation. This work is conducted in the framework of a medical evaluation campaign, which aims to evaluate and improve technologies to help patients and laypeople access eHealth data. To this end, the campaign is composed of different tasks, including a medical information retrieval (IR) task. Within this IR task, a web crawl of medically related documents, as well as patient queries are provided to participants. The queries are built to represent the potential information needs patients may have while reading their medical report. We start by describing typical types of patients' information needs. We then describe how these queries have been manually generated from medical reports for the first two years of the eHealth campaign. We then explore techniques that would enable us to automate the query generation process. This process is particularly challenging, as it requires an understanding of the patients' information needs, and of the electronic health records. We describe various approaches to automatically generate potential patient queries from medical reports and describe our future development and evaluation phase.

VisLR:
Visualization as Added Value in the Development, Use and Evaluation of Language Resources

31 May 2014

ABSTRACTS

Editors:

Annette Hautli-Janisz, Verena Lyding, Christian Rohrdantz

Workshop Programme

09:00 – 10:30 – Morning Session, Part I

09:00 – 09:10 – Introduction

09:10 – 09:40
Thomas Mayer, Johann-Mattis List, Anselm Terhalle and Matthias Urban, *An Interactive Visualization of Crosslinguistic Colexification Patterns*

09:40 – 10:00
Roberto Theron and Eveline Wandl-Vogt, *The Fun of Exploration: How to Access a Non-Standard Language Corpus Visually*

10:00 – 10:30
Pierrick Bruneau, Olivier Parisot, Amir Mohammadi, Cenk Demiroğlu, Mohammad Ghoniem and Thomas Tamisier, *Finding Relevant Features for Statistical Speech Synthesis Adaptation*

10:30 – 11:00 Coffee break

11:00 – 13:00 – Morning Session, Part II

11:00 – 11:30
Florian Stoffel, Dominik Jäckle and Daniel A. Keim, *Enhanced News-reading: Interactive and Visual Integration of Social Media Information*

11:30 – 11:50
Markus John, Florian Heimerl, Andreas Müller and Steffen Koch, *A Visual Focus+Context Approach for Text Comparison Tasks*

11:50 – 12:10
Miriam Butt, Tina Bögel, Kristina Kotcheva, Christin Schätzle, Christian Rohrdantz, Dominik Sacha, Nicole Dehe and Daniel A. Keim, *V1 in Icelandic: A Multifactorical Visualization of Historical Data*

12:10 – 12:40
Stefan Jänicke, Marco Büchler, Gerik Scheuermann, *Improving the Layout for Text Variant Graphs*

12:40 – 13:00 – Closing Discussion

Workshop Organizers

Annette Hautli-Janisz	University of Konstanz
Verena Lyding	European Academy of Bolzano/Bozen
Christian Rohrdantz	University of Konstanz

Workshop Programme Committee

Noah Bubenhofer	Dresden University of Technology
Miriam Butt	University of Konstanz
Chris Culy	University of Tübingen
Christopher Collins	University of Ontario Institute of Technology
Annette Hautli-Janisz	University of Konstanz
Gerhard Heyer	Leipzig University
Kris Heylen	University of Leuven
Daniel Keim	University of Konstanz
Steffen Koch	University of Stuttgart
Verena Lyding	European Academy of Bolzano/Bozen
Thomas Mayer	Philipps-Universität Marburg
Daniela Oelke	DIPF Frankfurt
Christian Rohrdantz	University of Konstanz

Preface to the VisLR Workshop

The VisLR workshop aims at bringing together people from visual analytics and computational linguistics to discuss the potentials and the challenges related to visualizing language data and in particular language resources. Linguistics has a long tradition of visually representing language patterns, from tree representations in syntax to spectograms in phonetics. However, the large amounts and ever-increasing complexity of today's resources call for new ways of visually encoding a multitude of abstract information on language in order to assure and enhance the quality and usability of these language resources.

We invited submissions on research demonstrating the development, use and evaluation of visualization techniques for language resources. This includes work applying existing visualization techniques to language resources as well as research on new visualization techniques that are specifically targeted to the needs of language resources.

The workshop contributions comprise visualization approaches for lexicographic data, text resources as well as speech data. Mayer et al. and Theron & Wandl-Vogt present two visualizations for facilitating the interactive exploration of lexicographic data. Bruneau et al. show how to make use of visual tools for analyzing high-dimensional models for speech synthesis adaptation. The visualizations for text and corpus data propose visual approaches that can be applied to support enhanced news-reading (Stoffel et al.), distant reading (John et al.) and the study of language change (Butt et al.) as well as the comparison of different editions of a text (Jänicke et al.).

Morning Session, Part I

Saturday 31 May, 9:00 – 10:30
Chairperson: Annette Hautli-Janisz

An Interactive Visualization of Crosslinguistic Colexification Patterns

Thomas Mayer, Johann-Mattis List, Anselm Terhalle and Matthias Urban

In this paper, we present an interactive web-based visualization for the CLICS database, an online resource for synchronic lexical associations (colexification patterns) in over 200 language varieties. The associations cover 1,288 concepts and represent the tendency for concepts to be expressed by the same words in the same languages and language varieties of the world. The complexity of the network structure in the CLICS database calls for a visualization component that makes it easier for researchers to explore the patterns of crosslinguistic colexifications. The network is represented as a force-directed graph and features a number of interactive components that allow the user to get an overview of the overall structure while at the same time providing an opportunity to look into the data in more detail. An integral part of the visualization is an interactive listing of all languages that contribute to the strength of a given pattern of colexification. Each language in the list is thereby attributed a different color depending on its genealogical or areal affiliation. In this way, given associations can be inspected for genealogical or areal bias.

The Fun of Exploration: How to Access a Non-Standard Language Corpus Visually

Roberto Theron and Eveline Wandl-Vogt

Historical dictionaries and non-standard language corpora can greatly benefit from a visual access in order to grasp the inherent tangled and complex nature of the knowledge encapsulated in them. Although visual analytics has been used to tackle a number of language and document related problems, most dictionaries are still reproducing the book metaphor in which Web pages substitute the paper and the user experience is only enhanced by means of hyperlinks. Although fields such as dialectology and dialectal lexicography have incorporated Geographic Information Systems and advanced computational linguistics features, spatio-temporal dynamics can be discovered or understood if appropriate visual analytics techniques are used to surpass the idea of these linguistic resources as alphabetically ordered lists. In this paper we present the work carried out in this direction for the Dictionary of Bavarian Dialects in Austria. By means of multiple-linked views an access that fosters the exploratory analysis of the data is enabled.

Finding Relevant Features for Statistical Speech Synthesis Adaptation

Pierrick Bruneau, Olivier Parisot, Amir Mohammadi, Cenk Demiroğlu, Mohammad Ghoniem and Thomas Tamisier

Statistical speech synthesis (SSS) models typically lie in a very high-dimensional space. They can be used to allow speech synthesis on digital devices, using only few sentences of input by the user. However, the adaptation algorithms of such weakly trained models suffer from the high dimensionality of the feature space. Because creating new voices is easy with the SSS approach, thousands of voices can be trained and a Nearest-Neighbor (NN) algorithm can be used to obtain better speaker similarity in those limited-data cases. NN methods require good distance measures that correlate well with human perception. This paper investigates the problem of finding good low-cost metrics, i.e. simple functions of feature values that map with objective signal quality metrics.

We show this is a ill-posed problem, and study its conversion to a tractable form. Tentative solutions are found using statistical analyzes. With a performance index improved by 36% w.r.t. a naive solution, while using only 0.77% of the respective amount of features, our results are promising. Deeper insights in our results are then unveiled using visual methods, namely high-dimensional data visualization and dimensionality reduction techniques. Perspectives on new adaptation algorithms, and tighter integration of data mining and visualization principles are eventually given.

Morning Session, Part II

Saturday 31 May, 11:00 – 13:00
Chairperson: Verena Lyding

Enhanced News-reading: Interactive and Visual Integration of Social Media Information

Florian Stoffel, Dominik Jäckle and Daniel A. Keim

Today, everyone has the possibility to acquire additional information sources as supplement to articles from newspapers or online news. The limitations of classical newspaper articles and restrictions of additional materials on online newsportals often lead to the situation where the reader demands additional news sources and more detailed information. When using the Internet, exploiting new information sources is a trivial task. Besides professionally administered information sources, like for example large newsportals such as cnn.com, there is a growing amount of user generated content available. Services like Twitter, Facebook or Reddit allow free discussion of any subject, giving everyone the possibility to participate. In this paper, we demonstrate an approach that combines professionally generated news content with user-generated data. This approach effectively enriches the information landscape and broadens the context of a given subject. For the presented system, we focus on Reddit, one of the biggest web portals for user-generated contents. Taking the general nature of user generated content into account, we exploit metadata and apply Natural Language Processing (NLP) methods to allow users to filter additional information, which is also supported visually.

A Visual Focus+Context Approach for Text Comparison Tasks

Markus John, Florian Heimerl, Andreas Müller and Steffen Koch

The concept of distant reading has become an important subject of digital humanities research. It describes a mode of textual work in which scholars are aided by automatic text analysis and visualization to directly find and access information relevant to their research questions in a large volume of text. While such techniques have proven to be effective in saving time and effort compared to extracting the information by linearly reading through the text, they introduce a new abstract level of analysis that masks the original source text. In this work, we present a flexible focus+context approach that facilitates scholarly textual work while at the same time supports efficient distant reading techniques. Users have full access to digital text sources for the perusal of a single text passage or the comparison of multiple ones. For each selected passage, an interactive visual summarization of its respective context allows users to effortlessly switch back and forth from close to distant reading. We demonstrate the capabilities of our approach with a usage scenario from the comparative study of poetics. The applicability and usefulness based on expert feedback is discussed afterwards.

V1 in Icelandic: A Multifactorical Visualization of Historical Data

Miriam Butt, Tina Bögel, Kristina Kotcheva, Christin Schätzle, Christian Rohrdantz, Dominik Sacha, Nicole Dehe and Daniel A. Keim

We present an innovative visualization technique for the analysis of historical data. We illustrate our method with respect to a diachronic case study involving V1 word order in Icelandic. A number of interacting factors have been proposed by linguists as being determinative of matrix declarative V1. The significance of these factors in contributing to declarative V1 can be explored interactively via our multifactorial visualization within a given text, but also comparatively over time. We apply the visualization to a corpus study based on the IcePaHC historical corpus of Icelandic and show that new results emerge very clearly out of the visualization component and that the appearance of declarative V1 is not confined to the situations identified so far by linguists. We demonstrate that the multifactorial visualization opens up new avenues for the exploration of alternative explanations. The visualization can be applied to any linguistic problem studying an interaction between several factors across time.

Improving the Layout for Text Variant Graphs

Stefan Jänicke, Marco Büchler, Gerik Scheuermann

Sentence Alignment Flows are visualizations for Text Variant Graphs that show the variations between different editions of texts. Although the resultant graph layouts are a substantial improvement to standard tools that are used in the corresponding Digital Humanities research field, the visualization is often cluttered due to large amounts of edge crossings and the occlusion of edges and vertices. In this paper, we present methods for the layering of vertices, the bundling of edges and the removal of overlaps between edges and vertices to reduce clutter, and therefore, to improve the readability for such graphs. Finally, we present the results of our survey with participants from the humanities and computer science, who had the task to compare the readability of Sentence Alignment Flows to the layouts generated by our improved method.

Challenges in the Management of Large Corpora (CMLC-2)

31 May 2014

ABSTRACTS

Editors:

Marc Kupietz, Hanno Biber, Harald Lüngen, Piotr Bánski, Evelyn Breiteneder, Karlheinz Mörth, Andreas Witt, Jani Takhsha

Workshop Programme

09:10 – 09:30 – Welcome and Introduction

09:10 – 09:30
Marc Kupietz, Harald Lüngen, Piotr Bański and Cyril Belica,
Maximizing the Potential of Very Large Corpora: 50 Years of Big Language Data at IDS Mannheim

9:30 – 10:00
Adam Kilgarriff, Pavel Rychlý and Miloš Jakubíček,
Effective Corpus Virtualization

10:00 – 10:30
Dirk Goldhahn, Steffen Remus, Uwe Quasthoff and Chris Biemann
Top-Level Domain Crawling for Producing Comprehensive Monolingual Corpora from the Web

10:30 – 11:00 Coffee break

11:00 – 11:30
Vincent Vandeghinste and Liesbeth Augustinus,
Making a large treebank searchable online. The SONAR case.

11:30 – 12:00
John Vidler, Andrew Scott, Paul Rayson, John Mariani and Laurence Anthony
Dealing With Big Data Outside Of The Cloud: GPU Accelerated Sort

12:00 – 12:30
Jordi Porta
From Several Hundred Million Words to Near One Thousand Million Words: Scaling Up a Corpus Indexer and a Search Engine with MapReduce

12:30 – 12:50
Hanno Biber and Evelyn Breiteneder
Text Corpora for Text Studies. About the foundations of the AAC-Austrian Academy Corpus

12:50 – 13:00 – Closing remarks

Workshop Organizers

Marc Kupietz	Institut für Deutsche Sprache, Mannheim
Hanno Biber	Institute for Corpus Linguistics and Text Technology, Vienna
Harald Lüngen	Institut für Deutsche Sprache, Mannheim
Piotr Bański	Institut für Deutsche Sprache, Mannheim
Evelyn Breiteneder	Institute for Corpus Linguistics and Text Technology, Vienna
Karlheinz Mörth	Institute for Corpus Linguistics and Text Technology, Vienna
Andreas Witt	Institut für Deutsche Sprache, Mannheim and University of Heidelberg

Workshop Programme Committee

Lars Borin	University of Gothenburg
Dan Cristea	"Alexandru Ioan Cuza" University of Iasi
Václav Cvrček	Charles University Prague
Mark Davies	Brigham Young University
Tomaž Erjavec	Jožef Stefan Institute, Ljubljana
Alexander Geyken	Berlin-Brandenburgische Akademie der Wissenschaften
Andrew Hardie	University of Lancaster
Nancy Ide	Vassar College
Miloš Jakubíček	Lexical Computing Ltd.
Adam Kilgarriff	Lexical Computing Ltd.
Krister Lindén	University of Helsinki
Jean-Luc Minel	Université Paris Ouest Nanterre La Défense
Christian Emil Ore	University of Oslo
Adam Przepiórkowski	Polish Academy of Sciences
Uwe Quasthoff	Leipzig University
Pavel Rychlý	Masaryk University Brno
Roland Schäfer	FU Berlin
Marko Tadić	University of Zagreb
Dan Tufiş	Romanian Academy, Bucharest
Tamás Váradi	Hungarian Academy of Sciences, Budapest

Introduction

We live in an age where the well-known maxim that “the only thing better than data is more data” is something that no longer sets unattainable goals. Creating extremely large corpora is no longer a challenge, given the proven methods that lie behind e.g. applying the Web-as-Corpus approach or utilizing Google's n-gram collection. Indeed, the challenge is now shifted towards dealing with large amounts of primary data and much larger amounts of annotation data. On the one hand, this challenge concerns finding new (corpus-)linguistic methodologies that can make use of such extremely large corpora e.g. in order to investigate rare phenomena involving multiple lexical items, to find and represent fine-grained sub-regularities, or to investigate variations within and across language domains; on the other hand, some fundamental technical methods and strategies are being called into question. These include e.g. successful curation of data, management of collections that span multiple volumes or that are distributed across several centres, methods to clean the data from non-linguistic intrusions or duplicates, as well as automatic annotation methods or innovative corpus architectures that maximise the usefulness of data or allow to search and to analyze it efficiently. Among the new tasks are also collaborative manual annotation and methods to manage it as well as new challenges to the statistical analysis of such data and metadata.

Maximizing the Potential of Very Large Corpora: 50 Years of Big Language Data at IDS Mannheim

Marc Kupietz, Harald Lüngen, Piotr Bański and Cyril Belica

Very large corpora have been built and used at the IDS since its foundation in 1964 and made available via the internet since the beginning of the 90's to currently over 30,000 researchers worldwide. The institute provides the largest archive of written German (Deutsches Referenzkorpus, DeReKo) which has recently been extended to 24 billion words. DeReKo has been managed and analysed by engines known as COSMAS and afterwards COSMAS II, which is currently being replaced by a new, scalable analysis platform called KorAP. KorAP makes it possible to manage and analyse texts that are accompanied by multiple, potentially conflicting, grammatical and structural annotation layers, and is able to handle resources that are distributed across different, and possibly geographically distant, storage systems. The majority of texts in DeReKo are not licensed for free redistribution, hence, the COSMAS and KorAP systems offer technical solutions to facilitate research on very large corpora that are not available (and not suitable) for download. For the new KorAP system, it is also planned to provide sandboxed environments to support non-remote-API access "near the data" through which users can run their own analysis programs.

Effective Corpus Virtualization

Adam Kilgarriff, Pavel Rychlý and Miloš Jakubíček

In this paper we describe an implementation of corpus virtualization within the Manatee corpus management system. Under corpus virtualization we understand logical manipulation with corpora or their parts grouping them into new (virtual) corpora. We discuss the motivation for such a setup in detail and show space and time efficiency of this approach evaluated on a 11 billion word corpus of Spanish.

Top-Level Domain Crawling for Producing Comprehensive Monolingual Corpora from the Web

Dirk Goldhahn, Steffen Remus, Uwe Quasthoff and Chris Biemann

This paper describes crawling and corpus processing in a distributed framework. We present new tools that build upon existing tools like Heritrix and Hadoop. Further, we propose a general workflow for harvesting, cleaning and processing web data from entire top-level domains in order to produce high-quality monolingual corpora using the least amount of language-specific data. We demonstrate the utility of the infrastructure by producing corpora for two under-resourced languages. Web corpus production for targeted languages and/or domains thus becomes feasible for anyone.

Making a Large Treebank Searchable Online. The SoNaR case.

Vincent Vandeghinste and Liesbeth Augustinus

We describe our efforts to scale up a syntactic search engine from a 1 million word treebank of written Dutch text to a treebank of 500 million words, without increasing the query time by a factor of 500. This is not a trivial task. We have adapted the architecture of the database in order to allow querying the syntactic annotation layer of the SoNaR corpus in reasonable time. We reduce the search space by splitting the data in many small databases, which each link similar syntactic patterns with sentence identifiers. By knowing on which databases we have to apply the XPath query we aim to reduce the query times.

Dealing with Big Data Outside of the Cloud: GPU Accelerated Sort

John Vidler, Andrew Scott, Paul Rayson, John Mariani and Laurence Anthony

The demands placed on systems to analyse corpus data increase with input size, and the traditional approaches to processing this data are increasingly having impractical run-times. We show that the use of desktop GPUs presents a significant opportunity to accelerate a number of stages in the normal corpus analysis pipeline. This paper contains our exploratory work and findings into applying high-performance computing technology and methods to the problem of sorting large numbers of concordance lines.

From Several Hundred Million to some Billion Words: Scaling Up a Corpus Indexer and a Search Engine with MapReduce

Jordi Porta

The growing size of corpora poses some technological challenges to their management. To reduce some of the problems arising in processing a few billion (10^9) words corpora, a shared-memory multithreaded version of MapReduce has been introduced into a corpus backend. Results on indexing very large corpora and computing basic statistics in this parallel processing framework on multicore computers are reported.

Text Corpora for Text Studies. About the foundations of the AAC-Austrian Academy Corpus.

Hanno Biber and Evelyn Breiteneder

One of the primary research aims of the research group at the "LIT – Literature in Transition" research group at the "Austrian Academy of Sciences" is to develop large text language resources for the study of literary texts. The paper will give an answer to the question how a significant historical text corpus of a considerable size that has been built along the principles of corpus

research can be made use of for practical research purposes in the field of literary text studies. The potential of a corpus-based methodological approach of literary text studies will be presented by investigating the textual qualities and the specific language use of literary texts. The paper will present several aspects of the research question illustrated by text examples and discuss the methodological implications of such a corpus based investigation thereby facing particular challenges with regard to large text corpora. Literary texts have properties that can be recognized and registered by means of a corpus-based study of language. Therefore the literary texts of the "AAC-Austrian Academy Corpus", a diachronic German language digital text corpus of more than 500 million tokens, will be used as examples for this research, thus constituting an ideal text basis for a corpus linguistic exploration into the fields of lexicographic units, syntactic constructions, narrative procedures and related issues concerned with the study of literary texts.

DIMPLE: DIsaster Management and Principled Large-scale information Extraction

31 May 2014

ABSTRACTS

Editors:

Khurshid Ahmad and Carl Vogel

Workshop Programme

INTRODUCTION

09:00-09:15 Khurshid Ahmad, Introduction: Disasters, Ethics, Terminology and Ontology

ONLINE INFORMATION SYSTEMS & DISASTER MANAGEMENT

09:15-09:35 Alexander Lörch & Mathias Bank, Topic Analyst® - A framework for recognizing early warnings

09:3-09:55 Enrico Musacchio & Francesco Russo, An Emergency Management System: Sistema Informativo Gestione Emergenze Protezione Civile

09:5-10:10 Questions & Discussion

COMMUNICATIONS DURING AND AFTER DISASTERS AND EMERGENCIES

10:1-10:30 Henrik Selsøe Sørensen, Multi-Lingual and Multi-Cultural Aspects of Post-Disaster Emergency Communication - the LinguaNet® Experience

10:30 - 11:00 Coffee break

11:00-11:20 Maria Teresa Musacchio, Social media and disaster management: US FEMA as a benchmark for its European counterparts?

11:20-12:00 Cilian Fennell, How Communications Companies Can Help Organisations Prepare for Disasters

1140-12:00 Maria Grazia Busa & Sara Brugnerotto, Italian doctor-patient interactions: a study of verbal and non-verbal behavior leading to miscommunication

12:00-12:15 Questions & Discussion

TRUST IN AND OF THE SOCIAL MEDIA

12:15-12:35 Carl Vogel, Anonymous FreeSpeech

13:35-12:55 Martin Mackin & Sadhbh McCarthy, Trust & Transparency in Social Media for Disaster Management

12:55-13:10 Questions & Discussion

13:10 - 14:00 Lunch

TERMINOLOGY AND ONTOLOGY OF DISASTERS

14:00-14:20 Bodil Nistrup Madsen & Hanne Erdman Thomsen, Terminological Ontologies for Risk and Vulnerability Analysis

14:20-14:40 Xiubo Zhang & Khurshid Ahmad, Ontology and terminology of disaster Management

14:40-15:00 Questions & Discussions

INFORMATION EXTRACTION FROM DISASTER DOCUMENT AND MEDIA STREAMS

15:00-15:20 Lars Döhling, Jirka Lewandowski & Ulf Leser, A Study in Domain-Independent Information Extraction for Disaster Management

15:20-15:40 Daniel Isemann, Andreas Niekler, Benedict Preler, Frank Viereck & Gerhard Heyer, OCR of Legacy Documents as a Building Block in Industrial Disaster Prevention

15:40-16:00 Stephen Kelly & Khurshid Ahmad, Determining levels of urgency and anxiety during a natural disaster: Noise, affect, and news in social media

16:00 - 16:30 Coffee break

16:30 - 17:00 Discussion

17:00 Workshop Closed

Workshop Organizers/Organizing Committee

Khurshid Ahmad	Trinity College Dublin, IRELAND.
Carl Vogel	Trinity College, Dublin, IRELAND.

Workshop Programme Committee

Khurshid Ahmad	Trinity College Dublin, IRELAND.
Gerhard Heyer	University of Leipzig, GERMANY
Linda Hogan	Trinity College, Dublin, IRELAND.
Bodil Madsen	Copenhagen Business School, DENMARK.
Sadhbh McCarthy	Centre for Irish and European Security, Dublin, IRELAND.
Maria Teresa Musacchio	University of Padova, ITALY.
Henrik Sørenson	Copenhagen Business School, DENMARK.
Carl Vogel	Trinity College, Dublin, IRELAND.

Disasters: Information management and Ethics

Saturday 31 May, 0900 – 1310

Chairperson: Khurshid Ahmad

Introduction: Disasters, Ethics, Terminology and Ontology

Khurshid Ahmad

This workshop addresses the use and implications of the use of technology in the management of the disaster life cycle: starting from warnings about impending disasters to suggestions about recovery. The technology issue has become more poignant with the advent of social media and its continually increasing use in disasters across the world. Consider the major disasters of this century, which has just begun, including hurricanes in the USA, earthquakes in Haiti, and tsunamis in Japan. In each of these cases there is documented evidence that social media is quite helpful in disseminating life and business critic al information quickly and effectively. The citizen is playing or should play an active role in monitoring, aver ting and rehabilitating before, during and after a disaster. One learns from the disaster archives referenced above that there is a need for an ethically well-grounded and accessible system that can harness the limitless data that streams through the social media and the formal media.

Topic Analyst® - A framework for recognizing early warnings

Alexander Lörch, Mathias Bank (a.loerch@cid.de, m.bank@cid.de)

The enormous quantity of information available today offers an unbelievable treasure of knowledge, which can provide relevant information to companies. The amount of data is however also a problem as it makes it very difficult to identify the relevant data and thus to generate knowledge. We present a comprehensive analysis tool named Topic Analyst® which enables the user to interactively investigate a huge amount of data. It enables the user to identify available topics, their trends and their tonality to make informed decisions.

An emergency management system: Sistema informativo gestione emergenze protezione civile

Enrico Musacchio and Francesco Russo
(enrico.musacchio@protecoeng.com;francescorusso@datapiano.it)

The software available today for managing civil protection actions during disasters emergencies is based on the collection of data about staff, work vehicles and rescue equipment, implemented and ordered in a structured database, used to raise the various resources, directing and coordinating them with human intervention. In many cases, the software is also able to display by means of interactions with a GIS, locating resources and means on the territory and possibly pre-established risk scenarios. In the light of current knowledge and awareness of the availability of modern facilities and web-based services that provide real-time interaction with the outside world, the described methodology of approach to the problem on which are based on the available software should be considered outdated and insufficient to ensure timely reaction and management of real-time communications. According to the developments of our research team, the approach to the problem of civil defense emergency management must be fully modernized, opening it to new technologies and services available, so that the reaction to adverse events can take place in real time and emergency operators can actually interact with the outside world and with social media, managing external communications and rescue operations with the necessary authority.

Multi-lingual and Multi-Cultural Aspects of Post-Disaster Emergency Communication the LinguaNet® Experience

Henrik Selsøe Sørensen (hss.ibc@cbs.dk)

LinguaNet® is a system for fast multi-lingual communications between police forces co-operating across frontiers. It has been in operation for more than two decades and proved its worth. From 1995 to 1998, CBS developed a number of add-ons to the system in the framework of an EU project in order to improve the multi-lingual and multi-cultural efficiency of the system. Three of these, namely multi-lingual casualty registration, cross-cultural ontology work, and a method for handling ontological discrepancies when country-specific data elements clash, are reported in this paper as examples. Given recent technological advances and the upcoming of social media as well as intelligent and instant big-data analyses since the reported research was carried out, it is suggested that the original ideas be revisited and re-engineered in view of improving efficiency in cross-frontier post-disaster emergency situations, where speed, robustness and reliability with respect to very divergent user profiles is a sine qua non.

Social media and disaster management: US FEMA as a benchmark for its European counterparts?

Maria Teresa Musacchio (mt.musacchio@unipd.it)

Over the last two decades disaster management and its public communication have been substantially transformed by the development of digital media such as blogs, wikis, social media and Youtube. Taken together, these shifts raise a series of issues for work in disaster management. Managing emergencies is a complex undertaking that relies extensively on knowledge management systems. Unlike European counterparts such as the German Bundesdienst für Bevolkerungsschutz und Katastrophenhilfe (BBK, Federal Office of Civil Protection and Disaster Assistance) and the Italian Protezione Civile (Civil Protection), the US Federal Emergency Management Agency (FEMA) employs social media technologies such as blogs, Facebook and Twitter as relevant disaster management and knowledge sharing mechanisms. This paper investigates how FEMA uses blogs, Facebook and Twitter, what knowledge is shared through these social media, and how knowledge sharing is facilitated and expedited through the use of these systems. Linguistic analysis is conducted to investigate the use of terminology, appraisal, syntactical structures, markers of shared and unshared information, thematic structure and text complexity with a view to identifying what knowledge is transmitted in emergencies, how alerts are provided that do not spread panic or descriptions supplied that form accountable reports of disaster management.

How Communications Companies Can Help Organisations Prepare for Disasters

Cilian Fennell (cilian@stillwater.ie)

Effective communication can play a crucial role in predicting, preventing and managing a disaster. A good communications company will help its clients prepare for crisis situations by building relationships with stakeholders, creating effective communications structures and developing a crisis communications strategy. For the purpose of this presentation, we will be concentrating on the case study of a flood management crisis in Ireland, a client project which we worked on recently.

Italian doctor-patient interactions: A study of verbal and non-verbal behavior leading to miscommunication

M. Grazia Busà, Sara Brugnerotto (mariagrazia.busa@unipd.it, sara.brugnerotto@studenti.unipd.it)

This study discusses aspects of doctor-patient communication and presents a preliminary analysis of doctor-patient interactions in Italy. The aim is to gain information on how (mis)communication between doctors and patients may affect the doctor-patient relationship and may lead patients to lack trust in their doctors. The authors use a corpus of existing audio-video materials on Italian doctor-patient interactions, and analyse doctors' use of verbal and nonverbal expressions in their exchanges with their patients. The analysis is aimed to identify which features may engender communication problems leading to misunderstandings and the perception of doctors as distant or unreliable. The preliminary findings reveal that patients' lack of trust in doctors may also be the result of doctors' use of culture-specific patterns of verbal and nonverbal expressions, for example certain sentences used for minimizing patients' fears, specific postures and gestures signalling distance or closure. These findings will be used for planning future investigations of doctor-patient interactions based on the collection and analysis of audio-visual material. Having a detailed knowledge of what patterns mostly affect communication in natural settings will provide important information to be implemented in digital devices.

Anonymous FreeSpeech

Carl Vogel (vogel@tcd.ie)

In public and private discourse, some may be heard to express disquiet about the supposed dangers of anonymity. Anonymous suggestion boxes may be classed with anonymous accusation of crime with the accusation forming the basis of legal proceedings. In some contexts, anonymity does appear to create danger. However, other contexts reveal that important benefits accrue from having the possibility of anonymous expression. Some of the literature on behavioral impacts of anonymity is reviewed with the aim of analyzing social media systems in light of their support for anonymous contribution. A new system is described. The new system supports anonymous communication, while thwarting some of the obvious risks that anonymity affords.

Trust in Social Media for Disaster Management

Sadhbh McCarthy, Martin Mackin (sadhbh@cies.ie)

Social media use in times of crisis and disaster has the potential to deliver many benefits. Uniquely, social media is a bilateral communication medium, which allows information to be conveyed and solicited between both the citizen and state. Social media is also a domesticated technology, which creates potential paths for surveillance crossover from the public into the private sphere. Any indications that state institutions (revenue commissioner, police, social welfare department) are abusing the technology - for example, by prying into citizens' behaviour - will have negative implications on the trust relationship between state and citizen. Crucially, it is this trust relationship that needs to be fostered, as those very state institutions will call upon it during times of crisis. Social media has also become deeply politicised, wherein parties, for their own ends, either valorise it as a societal innovation, an important component and contributor to the digital economy, or demonise it as the cause of societal ills (e.g. London riots, cyber bullying, harmful online fads) - of course, both are reductive, serving only to undermine and distort the true societal impact of social media. At its worst, social media can be can be seen as an engine of moral panic, occupying a space where headline driven journalism and opportunistic politics collide, which inevitability leads, over time, to further erosion of the bonds of trust between government stakeholders and the citizenry. In the absence of transparent, structured social media strategies, as part of a piece of coherent government policy and implemented at institutional level, the successful use of social media in disaster management (response and recovery) will be severely restricted. As it stands, government or party political interventions and engagement with social media tend to be focussed on narrow matters of voter mobilisation and engagement. While this approach can, in the short term, serve to deliver as part of a set of campaigning tools a temporary dividend in support, the cynical motives of it will, over time, only serve to undermine citizen trust. The power to harness the real benefits of social media will continue to be undermined as long as the societal impact of government constructs around social media do not recognize the important dynamics of such a trust relationship. A new paradigm of engagement, one that is rooted in trust, transparency and open dialogue, must be engendered before social media can fully realise its potential as a robust, reliable, effective communications component within crisis management strategies for periods of instability.

Disasters: Terminology, Ontology and Information Extraction

Saturday 31 May, 1400 – 1700

Chairperson: Gerhard Heyer

Terminological Ontologies for Risk and Vulnerability Analysis

Bodil Nistrup Madsen, Hanne Erdman Thomsen (bnm.ibc@cbs.dk, het.ibc@cbs.dk)

Risk and vulnerability analyses are an important preliminary stage in civil contingency planning. The Danish Emergency Management Agency has developed a generic model and a set of tools that may be used in the preparedness planning, i.e. for identifying and describing society's critical functions, for formulating threat scenarios and for assessing consequences. Terminological ontologies, which are systems of domain specific concepts comprising concept relations and characteristics, are useful, both when describing the central concepts of risk and vulnerability analysis (meta concepts), and for further structuring and enriching the taxonomies of society's critical functions and threats, which form an important part of the model. Creating terminological ontologies is a time consuming work, and therefore there is a need for automatic tools for extraction of terms, concept relations and characteristics. Terminological ontologies must adhere to a number of constraints, and therefore tools for automatic validation of these ontologies are also needed. Methods and tools for automatic ontology construction are being developed by researchers at Copenhagen Business School. The tools developed may also be used for extracting information on disasters from various media, and terminological ontologies may be used for enhancement of retrieval of information about disasters and for choosing the relevant countermeasures.

Ontology and Terminology of Disaster Management

Xiubo Zhang, Khurshid Ahmad (xizhang@scss.tcd.ie, kahmad@scss.tcd.ie)

The spate of natural disasters in the USA and in the European Union, especially floods and hurricanes, has led to the creation of specialised government agencies for dealing with such disasters in a unified command and control mode. Disaster management involves above all an inter-agency communication strategy that is at once transparent and deals in a timely manner with a set of unfolding events with the highest degree of professional care. Experts in various branches of engineering have to work together experts in health, in logistics, and in civil administration. Each has its own terminology generated from discipline-specific experiential knowledge. This experiential knowledge is highly codified for pedagogical purposes yet it is not accessible for the purposes of building information systems. We outline a corpus based method for building the ontology and terminology of natural disasters that relies in part on a legacy glossary of specific disasters and the structures and mnemonics used to encode the glossary are regarded as prototypical ontology of the disaster domain. Our methods and techniques were developed for looking at emerging specialised sciences, nanotechnology, breast cancer treatment, and more recently the inter-agency response to the 2008-financial debacle.

A Study in Domain-Independent Information Extraction for Disaster Management

Lars Döhling, Jirka Lewandowski, Ulf Leser (doehling, lewandow, leser@informatik.hu-berlin.de)

During and after natural disasters, detailed information about their impact is a key for successful relief operations. In the 21st century, such information can be found on the Web, traditionally provided by news agencies and recently through social media by affected people themselves. Manual information acquisition from such texts requires ongoing reading and analyzing, a costly process with very limited scalability. Automatic extraction offers fast information acquisition, but usually requires specifically trained extraction models based on annotated data. Due to changes in the language used, switching domains like from earthquake to flood requires training a new model in many approaches. Retraining in turn demands annotated data for the new domain. In this work, we study the cross-domain robustness of models for extracting casualty numbers from disaster reports. Our models are based on dictionaries, regular expressions, and patterns in dependency graphs. We provide an evaluation on extraction robustness across two disaster types earthquakes and floods. It shows that applying extra-domain models without retraining gives a relative F1 decrease of solely 9 %. This is a fairly small drop compared to previous results for similar complex extraction tasks.

OCR of Legacy Documents as a Building Block in Industrial Disaster Prevention

Daniel Isemann, Andreas Niekler, Benedict Preßler, Frank Viereck, Gerhard Heyer ({isemann, aniekler, heyerasv}@informatik.uni-leipzig.de, {mam10fdj, mam10hry}@studserv.uni-leipzig.de)

Legacy text documenting or recording operational details from industrial or other human-made facilities which constitute potential hazards may contain important safety critical information. In these cases it is desirable to make such safety critical information as is implicitly present in older records readily available and accessible to modern day authorities which have to deal with the facility from an industrial disaster prevention point of view. An important first step in such an analysis, we argue, is a robust optical character recognition (OCR) stage, for digitising often decade old records containing valuable information for industrial disaster prevention or management which may otherwise be lost or unavailable at the right time. In this paper we present an overview of a project concerned with the study and analysis of legacy records from the operational history of a deep geological repository for nuclear waste and present a preliminary study on segmenting letterhead information in legacy correspondence concerning this particular facility.

Determining levels of urgency and anxiety during a natural disaster: Noise, affect, and news in social media

Stephen Kelly, Khurshid Ahmad (kellys25@scss.tcd.ie, kahmad@scss.tcd.ie)

Since 2010, and perhaps before that as well, news and views of and about citizens caught up in a natural disaster, like floods and hurricanes, are increasingly available through digital media channels. In social media via Twitter for instance- and in formal media, especially in the blogs accompanying news compiled by various public and private sector agencies, one can get information about events as they unfold. Monitoring this stream of digital information provides valuable information for rescue agencies. However, caution has to be exercised in that this stream of information can be ostensibly stored for future analysis, by say resilience planners, without due care for the privacy of named entities, including individuals, places and institutions. In this paper we present a scalable bag-of-words method for analysing social media and crowd-sourced documents to visualise the evolving signature of a disaster event comprising disaster and affect terms. We illustrate our method by using a hurricane and an earthquake case study and two systems developed at Trinity College Dublin an ontology-based, scale-oriented system called Rocksteady and a terminology and ontology extraction system called CiCui. Ethical questions raised by automatic collection and analysis of social media data, especially the collation and storage of named entities are discussed.

LRE-REL2: 2nd Workshop on

Language Resources and Evaluation for Religious Texts

31 May 2014, Reykjavik, Iceland

A post-conference workshop co-hosted at LREC'2014 Language Resources and Evaluation Conference

ABSTRACTS

Editors:

Eric Atwell, Claire Brierley, Majdi Sawalha

Workshop Programme

31 May 2014

14:00 – 16:00 Session 1 Papers
14:00 Claire Brierley and Majdi Sawalha (Workshop Chairs)
Introduction to the 2nd Workshop on Language Resources and Evaluation for Religious Texts
14.10 Plenary Speaker: Majdi Sawalha (with co-authors Claire Brierley and Eric Atwell)
Automatically-generated, phonemic Arabic-IPA Pronunciation Tiers for the Boundary-Annotated Qur'an Dataset for Machine Learning (version 2.0)
14.40 Claudia Resch, Thierry Declerck, Barbara Krautgartner, and Ulrike Czeitschner
ABaC:us Revisited - Extracting and Linking Lexical Data from a Historical Corpus of Sacred Literature
14.50 Bruno Bisceglia, Rita Calabrese, Ljubica Leone
Combining Critical Discourse Analysis and NLP Tools in Investigations of Religious Prose
15.00 Daniela Gîfu, Liviu-Andrei Scutelnicu and Dan Cristea
Humour and Non-Humour in Religious Discourse
15.10 Manal AlMaayah, Majdi Sawalha and Mohammad A.M. Abushariah
A Proposed Model for Qur'anic Arabic Wordnet
15.20 Sameer M. Alrehaili and Eric Atwell
Computational Ontologies for Semantic Tagging of the Qur'an: A Survey of Past Approaches
15.30 Ahmad Alqurneh and Aida Mustapha
Traditional vs. Chronological Order: Stylistic Distance Analysis in Juz' Amma
15.40 Kamal Abou Mikhael
The Greek-Arabic New Testament Interlinear Process: greekarabicnt.org

16:00 – 16:30 Coffee break

16:30 – 17:15 Session 2: Posters, Discussion, and Networking
Poster presentations from all authors listed above

17:15 – 18:00 Session 3: Plenary Discussion led by Workshop Chairs
Topic: *Research agenda for LRE-Rel*

18:00 End of Workshop

Workshop Organizers

Claire Brierley	University of Leeds, UK
Majdi Sawalha	University of Jordan, Jordan
Eric Atwell	University of Leeds, UK
Bassam Hammo	University of Jordan, Jordan

Workshop Programme Committee

Muhammad A.M. Abushariah	Computer Information Systems, University of Jordan, Jordan
Eric Atwell	School of Computing, University of Leeds, UK
Claire Brierley	School of Computing, University of Leeds, UK
Liviu Dinu	Centre for Computational Linguistics, University of Bucharest, Romania
Kais Dukes	School of Computing, University of Leeds, UK
Moshe Koppel	Department of Linguistics, University of Jordan, Jordan
Dag Haug	Department of Philosophy, History of Art and Ideas, University of Oslo, Norway
John Lee	Halliday Centre for Intelligent Applications of Language Studies, City University of Hong Kong, (HK)
Deryle Lonsdale	Department of Linguistics and English Language, Brigham Young University, US
Bassam Hammo	Department of Computer Science, Bar-Ilan University, Israel
Bob MacDonald	Research and Development, Anthony Macauley Associates, Canada
Sane Yagi	Mathematics and Computer Science, Mohammed 1st University, Morocco
Mohamed Menacer	Taibah University, Saudi Arabia
Behrooz Minaei	School of Computer Engineering, Iran University of Science and Technology, Iran
Aida Mustapha	Department of Computer Science and Information Technology, Putra University, Malaysia
Nadeem Obaid	Computer Information Systems, University of Jordan, Jordan
Nils Reiter	Department of Computational Linguistics, Heidelberg University, Germany
Claudia Resch	Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences, Austria
Mortaza Rezaee	Islamic College, London, UK
Majdi Sawalha	Computer Information Systems, University of Jordan, Jordan
Gurpreet Singh	Centre for Language and Communication Studies, Trinity College Dublin, Ireland
Janet Watson	Arabic and Middle Eastern Studies, and Linguistics and Phonetics, University of Leeds, UK
Andrew Wilson	Department of Linguistics and English Language, University of Lancaster, UK
Azzeddine Mazroui	Computer Information Systems, University of Jordan, Jordan

Language Resources and Evaluation for Religious Texts 2: Introduction

Claire Brierley, Majdi Sawalha and Eric Atwell

Welcome to LRE-Rel2, the second *Workshop on Language Resources and Evaluation for Religious Texts*. After a successful launch at *LREC 2012* in Istanbul, Turkey, we have organised this second workshop hosted by *LREC 2014* in Reykjavik, Iceland. This is an inclusive workshop, aimed at researchers with a generic interest in religious texts to raise awareness of different perspectives and practices, and to identify some common themes. Our first workshop attracted a range of scholarship, particularly on Arabic and Islamic Studies, and this year we were keen to extend this range to canonical texts from other languages and religions, and to foster inter-faith corpus studies, tracing similarities as well as differences in religious texts, where this genre includes: the faith-defining religious canon; authoritative interpretations and commentary; sermons, liturgy, prayers, poetry, and lyrics.

We therefore welcomed submissions on a range of topics, including but not limited to:

- measuring semantic relatedness between multiple religious texts and corpora from different religions;
- analysis of ceremonial, liturgical, and ritual speech; recitation styles; speech decorum; discourse analysis for religious texts;
- formulaic language and multi-word expressions in religious texts;
- suitability of modal and other logic types for knowledge representation and inference in religious texts;
- issues in, and evaluation of, machine translation in religious texts;
- text-mining, stylometry, and authorship attribution for religious texts;
- corpus query languages and tools for exploring religious corpora;
- dictionaries, thesaurai, Wordnet, and ontologies for religious texts;
- new) corpora and rich and novel annotation schemes for religious texts;
- annotation and analysis of religious metaphor;
- genre analysis for religious texts;
- application in other disciplines e.g. theology, classics, philosophy, literature) of computer-mediated methods for analysing religious text

Our own research has focussed on Arabic Natural Language Processing, and in particular, *Qur'anic* Arabic (*cf.* our papers in both the *LRE-Rel2* Workshop and main *LREC 2014* Conference Proceedings); but we were pleased to receive papers dealing with a range of other holy books and religious texts, both historical and contemporary, with an interesting focus this year on the vernacular and on register *e.g.* historical German (1650-1750), and manifestations of humour in Romanian sermons. Many of the papers present an analysis technique applied to a specific religious text, which could also be relevant to analysis of other texts, including: automated, IPA-based transcription; specification of search patterns via regular expressions; stylometry, and detecting similarities and correspondences between texts; text extraction; semantic annotation and modelling; genre and critical discourse analysis. As an innovation this year, we will seek to identify a common research agenda for LRE-Rel during the plenary session.

This LRE-Rel Workshop demonstrates that religious texts are interesting and challenging for Language Resources and Evaluation researchers. It also shows LRE researchers a way to reach beyond our research community to the billions of readers of these holy books; LRE research can have a major impact on society, helping the general public to access and understand religious texts.

Session 1 Papers

Saturday 31 May 2014
Chairpersons: Claire Brierley and Majdi Sawalha

Automatically-generated, phonemic Arabic-IPA Pronunciation Tiers for the Boundary-Annotated Qur'an Dataset for Machine Learning (version 2.0)

Majdi Sawalha, Claire Brierley, and Eric Atwell

In this paper, we augment the *Boundary Annotated Qur'an* dataset published at LREC 2012 (Brierley *et al* 2012; Sawalha *et al* 2012a) with automatically generated phonemic transcriptions of Arabic words. We have developed and evaluated a comprehensive grapheme-phoneme mapping from Standard Arabic > IPA (Brierley *et al* [1]), and implemented the mapping in Arabic transcription technology which achieves 100% accuracy as measured against two gold standards: one for Qur'anic or Classical Arabic, and one for Modern Standard Arabic (Sawalha *et al* [1]). Our mapping algorithm has also been used to generate a pronunciation guide for a subset of Qur'anic words with heightened prosody (Brierley *et al* [2]). This is funded research under the EPSRC "*Working Together*" theme.

ABaC:us Revisited - Extracting and Linking Lexical Data from a Historical Corpus of Sacred Literature

Claudia Resch, Thierry Declerck, Barbara Krautgartner, and Ulrike Czeitschner

In this submission we describe results of work within the ABaC:us project dedicated to the extraction of lexical data from a corpus of sacred literature written in historical German language: All tokens occurring in the corpus have been semi-automatically mapped onto their corresponding lemmas in modern High German, which is a major achievement of the project. We are currently developing a RDF and SKOS based model for the extracted lexical data, in order to support their linking to corresponding senses in the Linked Open Data (LOD) framework, with a focus on religious themes.We describe first the achieved state of the extracted lexicon and then the actual state of our semantic model for the lexicon, with some examples of the code and of the linking to lexical senses available in the LOD.

Combining Critical Discourse Analysis and NLP Tools in Investigations of Religious Prose

Bruno Bisceglia, Rita Calabrese, Ljubica Leone

The present paper aims to investigate the discourse strategies adopted in selected samples of religious prose with the aim to: 1. Identify the main text functions which may result in differing subsections of the texts in terms of exhortation, exposition, narration and argumentation; 2. Compare the main principles and methodological procedures of Critical Discourse Analysis (CDA) with the corpus data. CDA explores the relationship between the exercise of power and discourse structuring by looking at the structures of power, authority and ideology which underlie the structures of speech and writing. To verify the above assumptions, a corpus of documents released by the Second Vatican Council (1962-65) was collected and automatically parsed by using the language analysis tools available at the Visual Interactive Syntax Learning (VISL) website. Along

with the primary corpus, a smaller control corpus of documents issued by the First Vatican Council in 1869-70 was created and annotated to provide a comparable basis to the analysis of the data. Following the automatic procedure of detection and extraction of information from a parsed corpus, we have matched corpus-based evidence and linguistic diagnostics (e.g. modality) generally used in CDA to determine the level of recursion and innovation, authority and liberality characterizing our data.

Humour and Non-Humour in Religious Discourse

Daniela Gîfu, Liviu-Andrei Scutelnicu and Dan Cristea

The paper describes a pilot study focusing on the exploration of humorous features in religious texts. Although religious discourses have a strong dogmatic tonality, some constructions contain visible humorous features, especially adapted to the audience expectations. Humour facilitates the apprehension of the religious discourse, evidencing not only the oratory dimension of the speaker but also helping the receptor to better perceive the message while also inducing affective effects. A corpus of preaches (which is contrasting with liturgical texts) is collected, in which humour is marked on Adjectival Noun Phrases. We propose a patter-based method for identifying humour in religious discourses in which patterns are lexicalised regular expressions of word categories. Using a religious lexicon, we classified Adjectival Noun Phrases in religious and non-religious. The study is meant to create a tool for automatic detection of humour in religious discourses. Automatically annotated corpora of preaches could become sources for further research that would reveal different valences in the religious texts.

A Proposed Model for Qur'anic Arabic Wordnet

Manal AlMaayah, Majdi Sawalha and Mohammad A.M. Abushariah

Most recent Arabic language computing research focuses on modern standard Arabic, but the classical Arabic of the Qur’an has been relatively unexplored, despite the importance of the Qur’an to Islam worldwide which can be used by both scholars and learners.This research work proposes to develop a WordNet for Qur’an by building semantic connections between words in order to achieve a better understanding of the meanings of the Qur’anic words using traditional Arabic dictionaries and a Qur’an ontology. The Qur’an corpus will be used as text and Boundary Annotated Qur’an Corpus (Brierley et al, 2012) will be used to explore the root and Part-of-Speech for each word and the word by word English translations. Traditional Arabic dictionaries will be used to find the Arabic meaning and derived words for each root in the Qur'anic Corpus. Then, these words and their meanings (Arabic, English) will be connected together through semantic relations. The achieved Qur’anic WordNet will provide an integrated semantic Qur’anic dictionary for the Arabic and English versions of the Qur'an.

Computational Ontologies for Semantic Tagging of the Qur'an: A Survey of Past Approaches

Sameer M. Alrehaili and Eric Atwell

Recent advances in Text Mining and Natural Language Processing have enabled the development of semantic analysis for religious text, available online for free. The availability of information is a key factor in knowledge acquisition. Sharing information is an important reason for developing an

ontology. This paper reports on a survey of recent Qur'an ontology research projects, comparing them in 9 criteria. We conclude that most of the ontologies built for the Qur'an are incomplete and/or focused in a limited specific domain. There is no clear consensus on the semantic annotation format, technology to be used, or how to verify or validate the results.

Traditional vs. Chronological Order: Stylistic Distance Analysis in Juz' Amma

Ahmad Alqurneh and Aida Mustapha

This paper analyzes the smoothness sequence in different texts in order to investigate whether texts that are close in time (which implies they are in chronological order) are stylistically similar or otherwise. We propose a spatial metaphor of distance with graphical representation to show similarity and dissimilarity between the texts. For this purpose, we choose surahs containing oaths from Juz' Amma because oath is the common topic between the surahs in this Juz' and work the surahs in a group of three. The analysis is performed in three parts; first is the closeness in terms of distance between the surahs, second is the ordering sequence among the surahs, and third is the closeness against the ordering sequence among the three surahs. The analysis showed that in general, while it is true for smoothness sequence to be based on the traditional closeness, it is not necessarily true for chronological closeness.

The Greek-Arabic New Testament Interlinear Process: greekarabicnt.org

Kamal Abou Mikhael

Greek-Arabic New Testament (NT) interlinears have yet to be published in a digital text format. Moreover, a corpus-based approach to the evaluation of Arabic translations of the NT has yet to be employed. The resource needed to actualize both of these goals is an aligned database of the source text and translation. This paper describes a process for creating such a database that combines automation and manual editing to create the database from which interlinears and reverse interlinears can be generated. The approach is text-based and leverages the text processing commands, editors, and scripting languages available in the *nix environment and uses freely available linguistic tools and NT texts. The paper also discusses issues of Arabic morphology and orthography that must be addressed in the preparation and alignment of the data. It is based on an initiative by greekarabicnt.org to align the Smith-Van Dyck Arabic Translation of the NT with its source text, Mill's Textus Receptus. At the time of writing, Chapters 1 and 2 of the First Epistle of John have been aligned, and interlinear and reverse-interlinear prototypes have been created in electronic format.