lrec 2012 istanbul

Eighth International Conference on Language Resources and Evaluation

LREC 2012 SATELLITE WORKSHOPS

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

May 21-22 & May 26-27, 2012

Istanbul Lütfi Kirdar Convention & Exhibition Centre
Istanbul, Turkey

WORKSHOP ABSTRACTS

Editors: Please refer to each single workshop list of editors.
Editorial Assistance by: Sara Goggi, Hélène Mazo

LREC 2012, EIGHTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION

Title: LREC 2012 Workshop Abstracts

Distributed by:
ELRA – European Language Resources Association
55-57, rue Brillat Savarin
75013 Paris
France

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

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

ISBN 978-2-9517408-7-7
EAN 9782951740877

TABLE OF CONTENTS

Workshop on Indian Language and Data: Resources and Evaluation

21 May 2012

ABSTRACTS

Editors:

Girish Nath Jha, Kalika Bali, Sobha L.

Workshop Programme

08:30-08:40 – Welcome by Workshop Chairs

08:40-08:55 – Inaugural Address by Mrs. Swarn Lata, Head, TDIL, Dept of IT, Govt of India

08:55-09:10 – Address by Dr. Khalid Choukri, ELDA CEO

0910-09:45 – Keynote Lecture by Prof Pushpak Bhattacharyya, Dept of CSE, IIT Bombay.

09:45-10:30 – Paper Session I
Chairperson: Sobha L

- Somnath Chandra, Swaran Lata and Swati Arora, *Standardization of POS Tag Set for Indian Languages based on XML Internationalization best practices guidelines*
- Ankush Gupta, *A Generic and Robust Algorithm for Paragraph Alignment and its Impact on Sentence Alignment in Parallel Corpora*
- Malarkodi C.S and Sobha Lalitha Devi, *A Deeper Look into Features for NE Resolution in Indian Languages*

10:30 – 11:00 Coffee break + Poster Session
Chairperson: Monojit Choudhury

- Akilandeswari A, Bakiyavathi T and Sobha Lalitha Devi, *'atu' Difficult Pronominal in Tamil*
- Subhash Chandra, *Restructuring of Painian Morphological Rules for Computer processing of Sanskrit Nominal Inflections*
- Praveen Dakwale, Himanshu Sharma and Dipti Misra Sharma, *Anaphora Annotation in Hindi Dependency TreeBank*
- H. Mamata Devi, *On the Development of Manipuri-Hindi Parallel Corpus*
- Madhav Gopal, *Annotating Bundeli Corpus Using the BIS POS Tagset*
- Madhav Gopal and Girish Nath Jha, *Developing Sanskrit Corpora Based on the National Standard: Issues and Challenges*
- Ajit Kumar and Vishal Goyal, *Practical Approach For Developing Hindi-Punjabi Parallel Corpus*
- Sachin Kumar, Girish Nath Jha and Sobha Lalitha Devi, *Challenges in Developing Named Entity Recognition System for Sanskrit*
- Swaran Lata and Swati Arora, *Exploratory Analysis of Punjabi Tones in relation to orthographic characters: A Case Study*
- Diwakar Mishra, Kalika Bali and Girish Nath Jha, *Grapheme-to-Phoneme converter for Sanskrit Speech Synthesis*
- Aparna Mukherjee, *Phonetic Dictionary for Indian English*
- Sibansu Mukhapadyay, Tirthankar Dasgupta and Anupam Basu, *Development of an Online Repository of Bangla Literary Texts and its Ontological Representation for Advance Search Options*
- Kumar Nripendra Pathak, *Challenges in Sanskrit-Hindi Adjective Mapping*
- Nikhil Priyatam Pattisapu, Srikanth Reddy Vadepally and Vasudeva Varma, *Hindi Web Page Collection tagged with Tourism Health and Miscellaneous*
- Arulmozi S, Balasubramanian G and Rajendran S, *Treatment of Tamil Deverbal Nouns in BIS Tagset*

- Gurpreet Singh, *Letter-to-Sound Rules for Gurmukhi Panjabi (Pa): First step towards Text-to-Speech for Gurmukhi*
- Silvia Staurengo, *TschwaneLex Suite (5.0.0.414) Software to Create Italian-Hindi and Hindi-Italian Terminological Database on Food, Nutrition, Biotechnologies and Safety on Nutrition: a Case Study.*

11:00 – 12:00 – Paper Session II
Chairperson: Kalika Bali

- Shahid Mushtaq Bhat and Richa Srishti, *Building Large Scale POS Annotated Corpus for Hindi & Urdu*
- Vijay Sundar Ram R, Bakiyavathi T, Sindhujagopalan R, Amudha K and Sobha Lalitha Devi, *Tamil Clause Boundary Identification: Annotation and Evaluation*
- Manjira Sinha, Tirthankar Dasgupta and Anupam Basu, *A Complex Network Analysis of Syllables in Bangla through SyllableNet*
- Pinkey Nainwani, *Blurring the demarcation between Machine Assisted Translation (MAT) and Machine Translation (MT): the case of English and Sindhi*

12:00-12:40 – Panel discussion on "*India and Europe - making a common cause in LTRs*"
Coordinator: Nicoletta Calzolari
Panelists - Kahlid Choukri, Joseph Mariani, Pushpak Bhattacharya, Swaran Lata, Monojit Choudhury, Zygmunt Vetulani, Dafydd Gibbon

12:40- 12:55 – Valedictory Address by Prof Nicoletta Calzolari, Director ILC-CNR, Italy

12:55-13:00 – 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 Lab India, Bangalore
A. G. Ramakrishnan	IISc Bangalore
Amba Kulkarni	University of Hyderabad
Dafydd Gibbon	Universitat Bielefeld, Germany
Dipti Mishra Sharma	IIIT, Hyderabad
Girish Nath Jha	Jawaharlal Nehru University, New Delhi
Joseph Mariani	LIMSI-CNRS, France
Kalika Bali	Microsoft Research Lab India, Bangalore
Khalid Choukri	ELRA, France
Monojit Choudhury	Microsoft Research Lab India, Bangalore
Nicoletta Calzolari	ILC-CNR, Pisa, Italy
Niladri Shekhar Dash	ISI Kolkata
Shivaji Bandhopadhyah	Jadavpur University, Kolkata
Sobha L.	AU-KBC Research Centre, Anna University
Soma Paul	IIIT, Hyderabad
Umamaheshwar Rao	University of Hyderabad

Introduction

WILDRE – the first ‘Workshop on Indian Language Data: Resources and Evaluation’ is being organized in Istanbul, Turkey on 21st May, 2012 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 WILDRE is

- 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 34 full papers received for review, we selected 24 for presentation in the workshop (7 for oral and 17 as posters).

Paper Session 1

Monday 21 May, 09:45-10:30
Chairperson: Sobha L

Standardization of POS Tag Set for Indian Languages based on XML Internationalization best practices guidelines

Somnath Chandra, Swaran Lata and Swati Arora

This paper presents a universal Parts of Speech (POS) tag set using W3C XML framework covering the major Indian Languages. The present work attempts to develop a common national framework for POS tag-set for Indian languages to enable a reusable and extendable architecture that would be useful for development of Web based Indian Language technologies such as Machine Translation, Cross-lingual Information Access and other Natural Language Processing technologies. The present POS tag schema has been developed for 13 Indian languages and being extended for all 22 constitutionally recognized Indian Languages. The POS schema has been developed using international standards e.g. metadata as per ISO 12620:1999, schema as per W3C XML internationalization guidelines and one to one mapping labels used 13 Indian languages.

A Generic and Robust Algorithm for Paragraph Alignment and its Impact on Sentence Alignment in Parallel Corpora

Ankush Gupta

In this paper, we describe an accurate, robust and language-independent algorithm to align paragraphs with their translations in a parallel bilingual corpus. The paragraph alignment is tested on 998 anchors (combination of 7 books) of English-Hindi language pair of Gyan- Nidhi corpus and achieved a precision of 86.86% and a recall of 82.03%. We describe the improvement in performance and automation of text alignment tasks by integrating our paragraph alignment algorithm in existing sentence aligner framework. This experiment carried out with 471 sentences on paragraph aligned parallel corpus, achieved a precision of 94.67% and a recall of 90.44%. Using our algorithm results in a significant improvement of 16.03% in Precision and 23.99% in Recall of aligned sentences as compared to when unaligned paragraphs are given as input to the sentence aligner.

A Deeper Look into Features for NE Resolution in Indian Languages

Malarkodi C.S and Sobha Lalitha Devi

Named Entity Recognition (NER) is the task of identifying and classifying proper nouns such as names of person, organization, location, etc. NER is tremendously used in various applications namely information extraction, question-answering, cross-lingual information access and query processing. Named Entity identification is mostly done for a specific domain and a particular language. In this paper, we developed NER for different Indian languages by using machine learning technique. Here we identify the common minimal features in English and various Indian languages of different language family (Bengali, Marathi, Punjabi, Tamil and Telugu). Further we used language dependent features to improve the system performance. The main goal of our task is to develop the NER with basic features and attain high performance. Conditional Random Fields (CRF) is used to build the training model for Indian languages. We trained CRF with few basic features and yielded encouraging results for both language generic and language specific system.

Poster Session

Monday 21 May, 10:30 – 11:00
Chairperson: Monojit Choudhury

‘atu’ Difficult Pronominal in Tamil

Akilandeswari A, Bakiyavathi T and Sobha Lalitha Devi

The paper presents a detailed analysis of 'atu' in Tamil, which is an equivalent of 'it' in English. ‘atu’ has many roles such as third person neuter pronoun, emphatic and as nominalizer. In this paper we are considering ‘atu’ in one particular construction, where the relative participle verb is suffixed with ‘atu’ in a multiple embedded sentence. In this form ‘atu’ can be anaphoric and non-anaphoric. In this paper we give a detailed analysis of ‘atu’ in the above construction. Using the analysis we identify the anaphoric and non-anaphoric ‘atu’ and also the antecedent of the anaphoric ‘atu’ using CRFs.

Restructuring of Paninian Morphological Rules for Computer processing of Sanskrit Nominal Inflections

Subhash Chandra

Sanskrit is the morphologically rich language and known as the mother of all Indian languages. Panini is known for his Sanskrit grammar, particularly in his formulation of the 3,959 rules of Sanskrit morphology, syntax and semantics described Ashtadhyayi (AD). AD is the foundational text of the grammatical branch of the Vedanga. The rules have been set out, much in the way of a mathematical function, to define the basic elements of the language including sentence structure, vowels, consonants, nouns, and verbs. The paper presents a developed Morphological Analyzer for Sanskrit Nominal Inflections which is based on restructuring of Paninian morphological rules. Restructuring of morphological rules is performed by reordering the Paninian morphological rules for computational purpose. Author proposed three steps for Sanskrit nominal morphological analysis. There are two parts of the system. One is recognition of Nominal Inflections from Sanskrit Texts and other is analysis of nominal inflections. The recognition of all nominal inflections is done in Sanskrit texts with help of relational databases. Then the nominal word is sliced into a sequence of morphemes and stem. Finally details information including stem, suffix, case and number of those nominal words are provided. Evaluation report shows that the accuracy of the system is 85%. It is also found that the system fail to provide 15% correct results where 5% errors are generated due to improper recognition of nominal words, 8% errors for incorrect morphological analysis of nominal words and rest 2% errors occur because of unsuccessful analysis. Online version of the system is available which runs on an Apache Server.

Anaphora Annotation in Hindi Dependency TreeBank

Praveen Dakwale, Himanshu Sharma and Dipti Misra Sharma

In this paper, we propose a scheme for anaphora annotation in Hindi Dependency Treebank. The goal is to identify and handle the challenges that arise in the annotation of reference relations in Hindi. Some of the anaphora annotation issues specific to Hindi like distribution of markable span, along with other problems like sequential annotation, representation format, multiple referents etc. are identified. The scheme hence incorporates some issue-specific characteristics to resolve them, among which the key characteristic is the head-modifier separation in referent selection. The utilization of the modifier-modified dependency relations inside a markable span provides for this head-modifier distinction. A part of the Hindi Dependency Treebank, of around 2500 sentences has been annotated with anaphoric relations and an inter-annotator study was carried out which shows a significant agreement over selection of the head referent. The current annotation is done for a limited set of pronominal categories and in the future, we intend to include other categories into the annotation work, as well.

On the Development of Manipuri-Hindi Parallel Corpus

H. Mamata Devi

A sentence aligned parallel corpus is a useful resource for Cross Lingual Information Retrieval, Machine Translation and Computational Linguistics. This paper describes the development of a sentence aligned Manipuri-Hindi parallel corpus and a Corpus Tool. The corpus contains 30,000 pairs of aligned Manipuri-Hindi sentences. The Manipuri sentences were manually translated to its corresponding Hindi sentences. The contents of the corpus were collected from different domain and sub domains so as to represent a balance corpus. The Corpus Tool consists of a Corpus Manager, a Statistical Text Analyzer and a Concordancer. It can work on both mono-lingual and multi-lingual corpus in two different data format viz ISCII and UTF8 and is adaptable to other Indian languages. This paper also describes the functionality and features of the Corpus Tool with the results generated by it from a book.

Annotating Bundeli Corpus Using the BIS POS Tagset

Madhav Gopal

Bundeli is an Indo-Aryan language, spoken mainly in the southern districts of Uttar Pradesh and northern districts of Madhya Pradesh. Despite having a large number of native speakers, the language terribly lacks language resources, in terms of corpus, language technology tools, guidelines, standards etc.; and this is partially because of its being a non-scheduled language of India and partially because of lack of an interested research community. Unlike Braj and Awadhi, Bundeli has never been a medium of literary expression, and consequently it lacks sufficient written texts. In this research an attempt is being made to develop its corpus and other computational resources to keep this language at a par with its other counterparts in the region and also to save this from possible extinction. Considering the fact that a digital corpus, after it was tagged at the POS level, could become an indispensable resource for various NLP tasks, machine learning, cognitive linguistics, comparative linguistics and theoretical linguistics, the development of annotated Bundeli corpus is unavoidable. Typologically it is close to Hindi (sparsely described as a variety of Hindi), but it is significantly different from Hindi. This paper introduces the scheme of corpus annotation for this less resourced language, using the BIS POS tagset, a standard tagset for tagging all the Indian languages. The creation of Bundeli corpus will also be discussed briefly.

Developing Sanskrit Corpora Based on the National Standard: Issues and Challenges

Madhav Gopal and Girish Nath Jha

This paper addresses the issue of the development of annotated corpus of Sanskrit using the Bureau of Indian Standards (BIS) POS tagset, a standard tagset designed for tagging Indian languages. The BIS POS tagset is a hierarchical tagset developed by the Bureau of Indian Standards committee. The development of Sanskrit corpus is going on by various research communities but standardised tagged data are no where available. The planned corpus is intended to put in public domain for research and academic purposes. Every effort is being made to make the data more and more useful from computational perspective. We will also discuss the issues and challenges that emerge during the POS tagging.

Practical Approach for Developing Hindi-Punjabi Parallel Corpus

Ajit Kumar and Vishal Goyal

The backbone of statistical analysis of any languages is the availability of very large corpus. We are working on Statistical Machine Translation System and require very large line-aligned parallel corpus. A number of parallel corpora do exist but due to copy right or other legal issues these are not shared by the developers. So we are developing our own Hindi-Punjabi sentence aligned parallel corpus. In this paper we are discussing the various approaches used by different researchers to develop monolingual and parallel corpora with their advantages and limitations and tools and techniques used by us in corpus development. We have automated some part of corpus development and rest of the work is being done manually. We are taking typed text from various sources and aligning it, where ever parallel documents are not available Hindi text is being translated into Punjabi text by using existing machine translation system. In this paper we discussed the dual approach applied by us in the development of Hindi-Punjabi line-aligned parallel corpus.

Challenges in Developing Named Entity Recognition System for Sanskrit

Sachin Kumar, Girish Nath Jha and Sobha Lalitha Devi

In this paper, we discuss several challenges in developing Named Entity Recognition (NER) system for Sanskrit. The paper also presents a broad framework for a Name Entity Tagset for Sanskrit (NETS), suitability and process-flow of the hybrid approach for Sanskrit NER system. The paper mainly focuses on the issues related to developing NER system for Indian languages especially Sanskrit. It also talks about intermediate results and its analysis based on a machine learning algorithm i.e. Conditional Random Field (CRF) applied on Pancatantra.

Exploratory Analysis of Punjabi Tones in relation to orthographic characters: A Case Study

Swaran Lata and Swati Arora

Punjabi is known tonal language of Indo-Aryan family with a very wide linguistic coverage across two countries. Bimodal one Male and one Female Data Repository and analysis for Indo-Aryan languages especially Punjabi is presently non-existent. Tone is the inherent feature of Punjabi due to the presence of 5 tonal characters. These Tonal Characters are represented by the corresponding aspirated or un-aspirated and voiced or unvoiced forms and also marked with high rising tone / ó / and low rising tone /ò/ on top of the accompanying vowel. Gender specific samples of recorded data from native speakers is being used for the analysis of Punjabi Tones in relation to Orthographic characters i.e. ਭ(bh) /p/ with a tone, ਧ (dh) /t/ with a tone, ਢ (dh)/ʈ/ with a tone, ਘ (gh)/k/ with a tone, and ਝ (Jh)/tʃ/ with a tone. These orthographic characters have lost their aspiration and have become tonal over a period of time. This analysis will help in the Speech Technology Research.

Grapheme-to-Phoneme converter for Sanskrit Speech Synthesis

Diwakar Mishra, Kalika Bali and Girish Nath Jha

The paper presents a Grapheme-to-Phoneme (G2P) converter as a module for Sanskrit speech synthesis. While Spoken Sanskrit is used in a limited specific context, and for general purpose by a fewer number of people (according to census of India data), the socio-cultural value of the language retains its significance in the modern Indian milieu. Most of its significance lies in the knowledge and other discourse of Sanskrit. Hence, accessibility to Sanskrit resources is of utmost importance in India, and also in the world, this paper, presents the development of a standalone G2P converter for Sanskrit based on the model developed by HP Labs India (and released through Local Language

Speech technology Initiative). The paper, also describes, the mapping of characters as well as specific rules that are necessary for a conversion of orthographic representation to a phonetic representation of Sanskrit. The authors will also demonstrate the system with the presentation.

Phonetic Dictionary for Indian English

Aparna Mukherjee

The paper presents the task of building an electronic phonetic dictionary for Indian English, with an aim to have a common source of English words, as they are pronounced in Indian English. The dictionary is a customized version of a pre-existing dictionary – the Carnegie Mellon University (CMU) Pronouncing Dictionary. The differences between pronunciation given in the North American English based CMU Dictionary and Indian English pronunciation, were identified and categorized. The identified categories and patterns were searched and replaced with the desired phonemes with the help of advanced regular expressions with back references. The data was edited manually as well, to avoid any error, since generalization of pattern in natural language is difficult. The phonetic dictionary was then used to generate English words in Devanagari script by following a mapping algorithm. The phonetic dictionary thus built for Indian English pronunciation can be further used to generate words in any Indian script.

Development of an Online Repository of Bangla Literary Texts and its Ontological Representation for Advance Search Options

Sibansu Mukhapadyay, Tirthankar Dasgupta and Anupam Basu

This paper presents the development of an online repository of Bangla literary texts written by eminent Bangla writers. This will allow large readers of Bangla language to access a huge storehouse for Bangla literary resources in Unicode to read them online. Such a large collection of Bangla literary texts will help numerous computational linguists to perform different interesting language related analysis and build linguistic applications. The paper also proposes the major role of ontological design to establish a knowledge sharing system within building a grand literary corpus of Bangla language. Finally, the paper presents the ontological representational schema for storing such a large collection of data that may help different search engines to retrieve the different metadata related to the author and the document efficiently.

Challenges in Sanskrit-Hindi Adjective Mapping

Kumar Nripendra Pathak

In this paper author is describing the adjective identification and in handling process theatrically for Sanskrit-Hindi Machine Translation (SHMT). In the diverse linguistic scenario in India, MT is needed to transfer the knowledge from one language to another. The overwhelming literary superiority of Sanskrit has attracted intellectuals worldwide and attempts to translate desired Sanskrit texts into other languages have been made since 17th century. Both the languages differ at various levels. Transferring Sanskrit linguistic features to Hindi is similarly challenging given the structural nuances that Hindi has developed in the course of its evolution. In order to provide comprehensible translation, an MT system should be capable of identifying an adjective and map it correctly into Hindi.

Hindi Web Page Collection tagged with Tourism Health and Miscellaneous

Nikhil Priyatam Pattisapu, Srikanth Reddy Vadepally and Vasudeva Varma

Web page classification has wide number of applications in the area of Information Retrieval. It is a crucial part in building domain specific search engines. Be it ’Google Scholar’ to search for scholarly articles or ’Google news’ to search for news articles, searching within a specific domain is a common practice. Sandhan is one such project which offers domain specific search for Tourism and Health domains across 10 different Indian Languages. Much of the accuracy of a web page classification algorithm depends on the data it gets trained on. The motivation behind this paper is to provide a proper set of guidelines to collect and store this data in an efficient and an error free way. The major contribution of this paper would be a Hindi web page collection manually classified into Tourism,Health and Miscellaneous.

Treatment of Tamil Deverbal Nouns in BIS Tagset

Arulmozi S, Balasubramanian G and Rajendran S

In Modern Tamil, the major Parts-Of-Speech (POS) categories are nouns, verbs, adjectives and adverbs. Of these, verbs take nominal suffixes such as –al, -tal, -kai, etc. to form “verbal nouns”. Formation of a verbal noun is a regular and productive process in Modern Tamil. In this paper we attempt to examine the place of verbal nouns in Tamil language as well as in the Bureau of Indian Standards (BIS) guidelines for POS annotation with reference to Tamil language. We present a brief introduction to the ILCI Corpus and Tamil language followed by a detailed study of the POS Tagsets for Tamil in general and verbal nouns in particular. The peculiar behaviour of verbal nouns which made them to be subcategorized under verb is discussed elaborately.

Letter-to-Sound Rules for Gurmukhi Panjabi (Pa): First step towards Text-to-Speech for Gurmukhi

Gurpreet Singh

This article presents the ongoing work to develop Letter-to-Sound rules for Guru Granth Sahib, the religious scripture of Sikh religion. The corpus forming the basis for development of the rules is taken from EMILLE corpora. Guru Granth Sahib is collection of hymns by founders of Sikh religion. After presenting an overview of Guru Granth Sahib and IPA representation in section 1 and Text-to-Speech in section 2, Letter-to-Sound rules developed will be presented in section 3. This paper will close with final discussion and future directions in section 4. The work presented stand at the development stage and no testing or experiment have so far been performed. The intention is to develop the Text-to-Speech for Punjabi language after developing it for limited set of language available in Guru Granth Sahib.

TschwaneLex Suite (5.0.0.414) Software to Create Italian-Hindi and Hindi-Italian Terminological Database on Food, Nutrition, Biotechnologies and Safety on Nutrition: a Case Study

Silvia Staurengo

Working with Tlex Suite to create an Italian-Hindi terminological database represents the concretization of Italian researchers’ dreams to join love for both countries, languages share knowledge with technicians, by one hand, and, by the other, extend it to non-technicians users. This research aim to develop communication without the English medium “forced step” – where it is possible - between workers on the field and, by the axiom on that food and nutrition convey socio-cultural implications and information as well, publicize unknown aspects of both countries to non-technicians. In few years we would like to launch on the market wide range of merchandises from papers to multimedia Lexicon; Machine-readable dictionary; pc', tablets', smart phone's applications etc. Further, we would like to create Speech corpora, too. Right now, we are submitting our work on progress paper as we have just started the research. On this we have put all our passion for selected subjects as well as initial problems faced working on the field, so we talk about of Case Study. We hope that humble and not so rich presentation would be useful, interesting for the audience.

Paper Session 2

Monday 21 May, 11:00 – 12:00
Chairperson: Kalika Bali

Building Large Scale POS Annotated Corpus for Hindi & Urdu

Shahid Mushtaq Bhat and Richa Srishti

Creation of annotated corpus is very essential for the technology development in natural languages. World languages can be divided into resource rich languages like European Languages and resource poor languages like Indian Languages. The former languages have enough technology at their disposal and it was possible due to the availability of large scale language resources while the latter have lagged behind only due to poor resource scenario. Though the work in this direction for ILs started very late as compared to their European counterpart, it is gaining momentum now-a-days in the form of various projects.

This paper is an attempt to discuss the ongoing work of developing 69.723K POS annotated corpus for Hindi & 66.488K POS annotated corpus for Urdu, using the BIS annotation standards. We didn't annotate corpus directly, either manually or automatically, rather we made a transition from LDC-IL annotation Scheme (based on ILPOST) to the contemporary BIS Scheme (inspired by ILMT). This transition resulted in afore mentioned quantum of annotated corpus as per BIS Standards.

Tamil Clause Boundary Identification: Annotation and Evaluation

Vijay Sundar Ram R, Bakiyavathi T, Sindhujagopalan R, Amudha K and Sobha Lalitha Devi

Clause boundary identification has a significant role in NLP applications. It has been used to improve the performance of different practical NLP systems. In this paper, we present the various types of clausal structures that exist in Tamil language. The clausal sentences from Newspapers, Novels and Tourism domains were collected and variations in the clausal structures across domains were analysed. Here we discuss about the annotation of tags used for various clauses and have focused on the Inter- annotator agreement. Inter- annotator agreement is the relative level of agreement between annotators. We have used kappa coefficient as the agreement statistic, which is the measure of inter annotator agreement. We also present the Automatic Clause Boundary Identification System developed using CRFs technique. We have evaluated and discussed on the performance of the system.

A Complex Network Analysis of Syllables in Bangla through SyllableNet

Manjira Sinha, Tirthankar Dasgupta and Anupam Basu

In this paper we present a development of a SyllableNet for Bangla language. Here, nodes of a network are the syllables and an edge between two syllables signifies that the two syllables have occurred within a same word. Number of times the two syllables occurred in a word reflects the edge weight of the graph. We use two different data sets viz. the online rabindra rachanabali from the web and the standard Bangla Banan Obhidhan, to perform the analysis of the network. Critical analysis of the syllabic network shows a low distance and a high clustering coefficient when compared with an associated Erdos–Renyi graph and with a random network with the same distribution of connectivity. Our comparison of network numeric with that of the Portuguese and Chinese syllabic networks reveals that despite having different origins, all of these networks have shown similar structural properties in terms of average path length, clustering coefficient and distribution of connectivity.

Blurring the demarcation between Machine Assisted Translation (MAT) and Machine Translation (MT): the case of English and Sindhi

Pinkey Nainwani

Dealing with divergences, at present, is a major concern of any Machine Translation (MT) system. This paper is an attempt to classify and analyze different types of translation divergences between English-Sindhi on the lines of Dorr (‘90, ‘93, and ‘94) and further suggests some strategies to handle divergences with rule-based methods. The paper also discusses some of the recent studies on classification and handling of divergences involving Indian languages (Gupta, ‘09, Sinha, ‘05).

First Workshop on Language Resources and Technologies for Turkic Languages

21 May 2012

ABSTRACTS

Workshop Organizers/Editors

Kemal Oflazer	Carnegie Mellon University - Qatar
Mehmed Özkan	Boğaziçi University
Mehmet Uğur Doğan	Tübitak-Bilgem
Hakan Erdoğan	Sabancı University
Dilek Hakkani-Tür	Microsoft
Yücel Bicil	Tübitak-Bilgem
İlknur Durgar El-Kahlout	Tübitak-Bilgem
Şeniz Demir	Tübitak-Bilgem
Alper Kanak	Tübitak-Bilgem

Workshop Programme

14:00 – 14:10 Welcome
14:10 – 15:10 Oral Session - I

- Cengiz Acartürk and Murat Perit Çakır, *Towards Building a Corpus of Turkish Referring Expressions*
- Arianna Bisazza and Roberto Gretter, *Building a Turkish ASR System with Minimal Resources*
- Francis Tyers, Jonathan North Washington, Ilnar Salimzyanov and Rustam Batalov, *A Prototype Machine Translation System for Tatar and Bashkir Based on Free/Open-Source Components*

15:10 – 15:30 Poster Presentations

- Işın Demirşahin, Ayışığı Sevdik-Çallı, Hale Ögel Balaban, Ruket Çakıcı and Deniz Zeyrek, *Turkish Discourse Bank: Ongoing Developments*
- Seza Doğruöz, *Analyzing Language Change in Syntax and Multiword Expressions: A Case Study of Turkish Spoken in the Netherlands*
- Atakan Kurt and Esma Fatma Bilgin, *The Outline of an Ottoman-to-Turkish Machine Transliteration System*
- Vít Baisa and Vít Suchomel, *Large Corpora For Turkic Languages and Unsupervised Morphological Analysis*
- Ayışığı B. Sevdik-Çallı, *Demonstrative Anaphora in Turkish: A Corpus Based Analysis*
- Alexandra V. Sheymovich and Anna V. Dybo, *Towards a Morphological Annotation of the Khakass Corpus*

15:30 – 16:30 Coffee Break & Poster Session

16:30 – 17:50 Oral Session - II

- Benjamin Mericli and Michael Bloodgood, *Annotating Cognates and Etymological Origin in Turkic Languages*
- Özkan Kılıç and Cem Bozşahin, *Semi-Supervised Morpheme Segmentation without Morphological Analysis*
- Şükriye Ruhi, Kerem Eryılmaz and M. Güneş C. Acar, *A Platform for Creating Multimodal and Multilingual Spoken Corpora for Turkic Languages: Insights from the Spoken Turkish Corpus*
- Eray Yıldız and A. Cüneyd Tantuğ, *Evaluation of Sentence Alignment Methods for English-Turkish Parallel Texts*

17:50 – 18:00 Closing

Workshop Programme Committee

Yeşim Aksan	Mersin University
Adil Alpkoçak	Dokuz Eylül University
Mehmet Fatih Amasyalı	Yıldız Technical University
Ebru Arısoy	IBM T.J. Watson Research Center
Levent Arslan	Boğaziçi University
Barış Bozkurt	Bahçeşehir University
Cem Bozşahin	Middle East Technical University
Ruket Çakıcı	Middle East Technical University
Özlem Çetinoğlu	University of Stuttgart
Cemil Demir	Tübitak-Bilgem
Cenk Demiroğlu	Özyeğin University
Banu Diri	Yıldız Technical University
Gülşen Cebiroğlu Eryiğit	İstanbul Technical University
Engin Erzin	Koç University
Tunga Güngör	Boğaziçi University
Ümit Güz	Işık University
Yusuf Ziya Işık	Tübitak-Bilgem
Selçuk Köprü	Teknoloji Yazılımevi
Atakan Kurt	Fatih University
Oğuzhan Külekçi	Tübitak-Bilgem
Coşkun Mermer	Tübitak-Bilgem
Arzucan Özgür	Boğaziçi University
Fatma Canan Pembe	Tübitak-Bilgem
Şükriye Ruhi	Middle East Technical University
Murat Saraçlar	Google - Boğaziçi University
Bilge Say	Middle East Technical University
Ahmet Cüneyd Tantuğ	İstanbul Technical University
Erdem Ünal	Tübitak-Bilgem
Deniz Yüret	Koç University
Deniz Zeyrek	Middle East Technical University

Introduction

Turkic languages are spoken as a native language by more than 150 million people all around the world (one of the 15 most widely spoken first languages). Prominent members of this family are Turkish, Azerbaijani, Turkmen, Kazakh, Uzbek, and Kyrgyz. Turkic languages have complex agglutinative morphology with very productive inflectional and derivational processes leading to a very large vocabulary size. They also have a very free constituent order with almost no formal constraints. Furthermore, due to various historical and social reasons these languages have employed a wide-variety of writing systems and still do so. These aspects bring numerous challenges (e.g., data sparseness and high number of out-of-vocabulary words) to computational processing of these languages in tasks such as language modeling, parsing, statistical machine translation, speech-to-speech translation, etc. Thus, pursuing high-quality research in this language family is particularly challenging and laborious.

This workshop is timely as there is burgeoning interest in the field of research. Moreover, various language resources and computational processing techniques for Turkic languages need to be developed in order to bring their status up to par with more studied languages in the context of speech and language processing. It has become more crucial as the number of international affairs, economic activities, and cultural relations between Turkic people and EMEA (Europe, Middle East, and Africa) increase. There exist a growing demand and awareness on related research and current developments provide us with solutions from different approaches. However, there still remain many problems to be solved and much work to be done in the roadmap for Turkic languages.

The workshop will bring together the academicians, experts, research-oriented enterprises (SMEs, large companies, and potential end users), and all other stakeholders who are actively involved in the field of speech and language technologies for Turkic languages. The workshop will focus on cut-edge research and promote discussions to better disseminate knowledge and visionary thoughts for speech and language technologies aligned with Turkic languages. The workshop is expected to properly portray the current status of Turkic speech and language research performances, and to enlighten the pros and cons, end user needs, current state-of-the-art, and existing R&D policies and trend. This workshop will also have a positive impact on establishing a research community moving into the future and on building a collaboration environment which we anticipate to receive widespread attention in the HLT domain.

The workshop features 7 oral and 6 poster presentations. The accepted papers range from annotation initiatives to language and speech resources and technologies.

Oral Session - I
Monday 21 May, 14:10 – 15:10
Chairperson: Deniz Zeyrek

Towards Building a Corpus of Turkish Referring Expressions

Cengiz Acartürk and Murat Perit Çakır

In this paper we report on the preliminary findings of our ongoing study on Turkish referring expressions used in situated dialogs. Situated dialogs of pairs of Turkish speakers were collected while they were engaged with a collaborative Tangram puzzle solving task, which was designed by Spanger et al (2011) in an effort to build a corpus of referring expressions in Japanese and English. The paper provides our preliminary results on the Turkish corpus and compares them with the findings of comparable studies conducted on Japanese and English referring expressions.

Building a Turkish ASR system with minimal resources

Arianna Bisazza and Roberto Gretter

We present an open-vocabulary Turkish news transcription system built with almost no language-specific resources. Our acoustic models are bootstrapped from those of a well trained source language (Italian), without using any Turkish transcribed data. For language modeling, we apply unsupervised word segmentation induced with a state-of-the-art technique (Creutz and Lagus, 2005) and we introduce a novel method to lexicalize suffixes and to recover their surface form in context without need of a morphological analyzer. Encouraging results obtained on a small test set are presented and discussed.

A prototype machine translation system for Tatar and Bashkir based on free/open-source components

Francis Tyers, Jonathan North Washington, Ilnar Salimzyanov and Rustam Batalov

This paper presents a prototype bidirectional machine translation system between Tatar and Bashkir, two minority Turkic languages of Russia. While the system has low open-domain coverage, results are presented that suggest that high accuracy may be obtained between these two closely-related languages, on a par with similar systems.

Poster Presentations
Monday 21 May, 15:10 – 15:30
Chairperson: Ahmet Cüneyd Tantuğ

Turkish Discourse Bank: Ongoing Developments

Işın Demirşahin, Ayışığı Sevdik-Çallı, Hale Ögel Balaban, Ruket Çakıcı and Deniz Zeyrek

This paper describes the first release of the Turkish Discourse Bank (the TDB), the first large-scale, publicly available language resource with discourse-level annotations for Turkish. The TDB consists of a sub-corpus of the METU Turkish Corpus (MTC), which is annotated for discourse connectives; their arguments, i.e., the text spans they bring together; modifiers of the connectives, and supplementary spans that provide details for the arguments. In this paper, we describe the features of the MTC and the sub-corpus on which the TDB is built. We provide information about the annotations and other contents of the first release of the TDB. Finally, we describe the ongoing developments including annotating the sense and the class of the connectives, and the morphological features of the nominalized arguments of subordinating conjunctives.

Analyzing language change in syntax and multiword expressions: A case study of Turkish Spoken in the Netherlands

A. Seza Doğruöz

All languages change and spoken corpora provide opportunities to analyze linguistic changes while they are still taking place. Turkish spoken in the Netherlands (NL-Turkish) has been in contact with Dutch for over fifty years and it sounds different in comparison to Turkish spoken in Turkey (TR-Turkish). Comparative analyses of NL-Turkish and TR-Turkish spoken corpora do not reveal significant on-going changes in terms of word order. However, Dutch-like multiword expressions make NL-Turkish sound unconventional to TR-Turkish speakers. In addition to presenting these on-going changes, this study also discusses the challenges with respect to syntactic parsing as well as identification and classification of multiword expressions in spoken Turkish corpora.

The Outline of an Ottoman-to-Turkish Machine Transliteration System

Atakan Kurt and Esma Fatma Bilgin

The Ottoman script is a writing system of the Turkish language which was in use from the early the 13th century until the 20th century. The transliteration of Ottoman script to Latin-based modern Turkish script is necessary in order to make a huge collection of documents available to readers. The transliteration problem can be reduced to pronunciation generation in Turkish for the Ottoman script, because the pronunciation of words remains the same. The main problem of the transliteration is the lack of a regular of orthography in the Ottoman script. The complexity of the problem requires a combination of NLP techniques beyond simple character mappings. This paper outlines the Ottoman orthography in general and discusses the complexities, problems, difficulties, exceptional cases in the Ottoman orthography. Then the vowel and consonant mappings between the two scripts are defined. Finally we present the outline of an automatic machine transliteration framework from Ottoman to Turkish.

Large Corpora for Turkic Languages and Unsupervised Morphological Analysis

Vít Baisa and Vít Suchomel

In this article we describe six new web corpora for Turkish, Azerbaijani, Kazakh, Turkmen, Kyrgyz and Uzbek languages. The data for these corpora was automatically crawled from the web by SpiderLing. Only minimal knowledge of these languages was required to obtain the data in raw form. Corpora are tokenized only since morphological analyzers and disambiguators for these languages are not available (except for Turkish). Subsequent experiment with unsupervised morphological segmentation was carried out on the Turkish corpus. In this experiment we achieved encouraging results. We used data provided for MorphoChallenge competition for the purpose of evaluation.

Demonstrative Anaphora in Turkish: A Corpus Based Analysis

Ayışığı B. Sevdik-Çallı

This study investigates Turkish demonstrative anaphora including bare demonstrative uses and demonstrative NP uses on a 20K subpart of the METU Turkish Discourse Bank. Antecedents of demonstrative anaphora including abstract object references are identified in 10 texts of approximately 2000 words each. Preliminary analysis shows that for endophoric cases, where the antecedent of the anaphora can be identified in the text, references to concrete object antecedents of the three Turkish demonstratives *bu* ('this), *şu* ('this/that), and *o* ('that) is overall higher than references to abstract object antecedents. However, for the demonstrative *bu* ('this), abstract object reference is almost equal to the concrete object references. The antecedents of the third person singular pronoun *o* ('he/she/it) have also been identified as it is lexically identical to the distal demonstrative *o* ('that), and it was seen that the demonstrative pronoun use occurred as much as the personal pronoun use. The implications of these findings are discussed in terms of Turkish anaphora resolution.

Towards a Morphological Annotation of the Khakass Corpus

Alexandra V. Sheymovich and Anna V. Dybo

This paper describes development of a corpus of the Khakass language and design of a morphological parser for it. Being one of the RAS projects, it follows the RAS program in regard to the development of corpora for languages of the Russian Federation, including Turkic minority languages such as Khakass. Khakass is a language spoken by about 20,000 people, most of whom are bilingual in Russian. They live in the southern Siberian Khakass Republic in Russia. We present the preliminary linguistic work for creating automatic morphological annotation for the Khakass written corpus. Main components of this work are: 1) the database of the Khakass word stems generated by StarLing system, 2) the computational model of a Khakass wordform and 3) the set of phonetic rules that constrain the choice of allomorphs within the wordform. We also present Khakass inflectional affixes with their allomorphs.

Oral Session - II
Monday 21 May, 16:30 – 17:50
Chairperson: Kemal Oflazer

Annotating Cognates and Etymological Origin in Turkic Languages

Benjamin Mericli and Michael Bloodgood

Turkic languages exhibit extensive and diverse etymological relationships among lexical items. These relationships make the Turkic languages promising for exploring automated translation lexicon induction by leveraging cognate and other etymological information. However, due to the extent and diversity of the types of relationships between words, it is not clear how to annotate such information. In this paper, we present a methodology for annotating cognates and etymological origin in Turkic languages. Our method strives to balance the amount of research effort the annotator expends with the utility of the annotations for supporting research on improving automated translation lexicon induction.

Semi-supervised morpheme segmentation without morphological analysis

Özkan Kılıç and Cem Bozşahin

The premise of unsupervised statistical learning methods lies in a cognitively very plausible assumption that learning starts with an unlabeled dataset. Unfortunately such datasets do not offer scalable performance without some semi-supervision. We use 0.25% of METU-Turkish Corpus for manual segmentation to extract the set of morphemes (and morphs) in its 2 million word database without morphological analysis. Unsupervised segmentations suffer from problems such as oversegmentation of roots and erroneous segmentation of affixes. Our supervision phase first collects information about average root length from a small fragment of the database (5,010 words), then it suggests adjustments to structure learned without supervision, before and after a statistically approximated root, in an HMM+Viterbi unsupervised model of n-grams. The baseline of .59 f-measure goes up to .79 with just these two adjustments. Our data is publicly available, and we suggest some avenues for further research.

A Platform for Creating Multimodal and Multilingual Spoken Corpora for Turkic Languages: Insights from the Spoken Turkish Corpus

Şükriye Ruhi, Kerem Eryılmaz and M. Güneş C. Acar

Based on insights gained from the corpus design and corpus management work involved in the compilation of the Spoken Turkish Corpus (STC), this paper addresses the possibility of developing sustainable, comparable, multimodal spoken corpora for facilitating comparative studies on Turkic Languages, with the capacities of a digital platform that incorporates EXMARaLDA software suite and a web-based corpus management system (STC-CMS), which together provide an interoperable system that can be customized for the creation of spoken and written corpora. Section 2highlights the significance of multimodal corpus resources for comparative research and the development of technologies, and describes the implementation in STC, especially focusing on its metadata parameters and the flexibility of its transcription tools for representing cross-linguistic variation. Section 3 addresses the issue of developing common infrastructure for corpus compilation that can facilitate data transfer between resources. The paper concludes with a brief discussion on the challenge for creating comparable spoken corpora for the Turkic languages in regard to orthographic systems.

Evaluation of Sentence Alignment Methods for English-Turkish Parallel Texts

Eray Yıldız and A. Cüneyd Tantuğ

In this work, we evaluate the performances of sentence alignment methods on aligning English-Turkish parallel texts. Three publicly available tools employing different strategies are tested in our study: a sentence length-based alignment method, a lexicon-based alignment method and a machine translation based alignment method. Experiments are carried out on a test dataset of parallel texts collected from web, mostly from newspapers. Due to the highly inflectional and derivational morphological structure of Turkish, we have incorporated stemming pre-processing step for the lexicon based tests. However, finding stems of Turkish wordforms requires a full morphological analysis and morphological disambiguation. So, as a simpler alternative stemming method, we suggest taking only *k*-characters of the wordforms as stems. Our experiments show that lexicon based methods with stemming performs best among all methods.

Best Practices for Speech Corpora in Linguistic Research

21 May 2012

ABSTRACTS

Editors:

Michael Haugh, Sukrie Ruhi, Thomas Schmidt, Kai Wörner

Workshop Programme

14:00 – Case Studies: Corpora & Methods

Janne Bondi Johannessen, Øystein Alexander Vangsnes, Joel Priestley and Kristin Hagen:
A linguistics-based speech corpus

Adriana Slavcheva and Cordula Meißner:
GeWiss – a comparable corpus of academic German, English and Polish

Elena Grishina, Svetlana Savchuk and Dmitry Sichinava:
Multimodal Parallel Russian Corpus (MultiPARC): Multimodal Parallel Russian Corpus (MultiPARC): Main Tasks and General Structure

Sukriye Ruhi and E. Eda Isik Tas:
Constructing General and Dialectal Corpora for Language Variation Research: Two Case Studies from Turkish

Theodossia-Soula Pavlidou:
The Corpus of Spoken Greek: goals, challenges, perspectives

Ines Rehbein, Sören Schalowski and Heike Wiese:
Annotating spoken language

Seongsook Choi and Keith Richards:
Turn-taking in interdisciplinary scientific research meetings: Using 'R' for a simple and flexible tagging system

16:30 – 17:00 Coffee break

17:00 – Panel: Best Practices

Pavel Skrelin and Daniil Kocharov
Russian Speech Corpora Framework for Linguistic Purposes

Peter M. Fischer and Andreas Witt
Developing Solutions for Long-Term Archiving of Spoken Language Data at the Institut für Deutsche Sprache

Christoph Draxler
Using a Global Corpus Data Model for Linguistic and Phonetic Research

Brian MacWhinney, Leonid Spektor, Franklin Chen and Yvan Rose
Best Practices in the TalkBank Framework

Christopher Cieri and Malcah Yaeger-Dror
Toward the Harmonization of Metadata Practice for Spoken Languages Resources

Sebastian Drude, Daan Broeder, Peter Wittenburg and Han Sloetjes
Best practices in the design, creation and dissemination of speech corpora at The Language Archive

18:30 Final Discussion

Workshop Organizers

Michael Haugh	Griffith University, Australia
Sukryie Ruhi	Middle Easter Technical University, Ankara
Thomas Schmidt	Institut für Deutsche Sprache, Mannheim
Kai Wörner	Hamburger Zentrum für Sprachkorpora

Workshop Programme Committee

Yeşim Aksan	Mersin University
Dawn Archer	University of Central Lancashire
Steve Cassidy	Macquarie University, Sydney
Chris Christie	Loughborough University
Arnulf Deppermann	Institute for the German Language, Mannheim
Ulrike Gut	University of Münster
Iris Hendrickx	Linguistics Center of the University of Lisboa
Alper Kanak	Turkish Science and Technology Institute – TÜBİTAK
Kemal Oflazer	Carnegie Mellon at Qatar
Antonio Pareja-Lora	ILSA-UCM / ATLAS-UNED
Petr Pořízka	Univerzita Palackého
Jesus Romero-Trillo	Universidad Autonoma de Madrid
Yvan Rose	Memorial University of Newfoundland
Martina Schrader-Kniffki	University of Bremen
Deniz Zeyrek	Middle East Technical University

Call for Papers

This half-day-workshop addresses the question of best practices for the design, creation and dissemination of speech corpora in linguistic disciplines like conversation analysis, dialectology, sociolinguistics, pragmatics and discourse analysis. The aim is to take stock of current initiatives, see how their approaches to speech data processing differ or overlap, and find out where and how a potential for coordination of efforts and standardisation exists.

Largely in parallel to the speech technology community, linguists from such diverse fields as conversation analysis, dialectology, sociolinguistics, pragmatics and discourse analysis have, in the last ten years or so, intensified their efforts to build up (or curate) larger collections of spoken language data. Undoubtedly, methods, tools, standards and workflows developed for corpora used in speech technology often serve as a starting point and a source of inspiration for the practices evolving in the linguistic research community. Conversely, the spoken language corpora developed for linguistic research can certainly also be valuable for the development or evaluation of speech technology. Yet it would be an oversimplification to say that speech technology data and spoken language data in linguistic research are merely two variants of the same category of language resources. Too distinct are the scholarly traditions, the research interests and the institutional circumstances that determine the designs of the respective corpora and the practices chosen to build, use and disseminate the resulting data.

The aim of this workshop is therefore to look at speech corpora from a decidedly linguistic perspective. We want to bring together linguists, tool developers and corpus specialists who develop and work with authentic spoken language corpora and discuss their different approaches to corpus design, transcription and annotation, metadata management and data dissemination. A desirable outcome of the workshop would be a better understanding of

- best practices for speech corpora in conversation analysis, dialectology, sociolinguistics, pragmatics and discourse analysis,
- possible routes to standardising data models, formats and workflows for spoken language data in linguistic research
- ways of linking up trends in speech technology corpora with corresponding work in the linguistics communities

Topics of interest include:

- speech corpus designs and corpus stratification schemes
- metadata descriptions of speakers and communications
- legal issues in creating, using and publishing speech corpora for linguistic research
- transcription and annotation tools for authentic speech data
- use of automatic methods for tagging, annotating authentic speech data
- transcription conventions in conversation analysis, dialectology, sociolinguistics, pragmatics and discourse analysis
- corpus management systems for speech corpora
- workflows and processing chains for speech corpora in linguistic research
- data models and data formats for transcription and annotation data
- standardization issues for speech corpora in linguistic research
- dissemination platforms for speech corpora
- integration of speech corpora from linguistic research into digital infrastructures

Case Studies: Corpora & Methods

Monday 21 May, 14:00-16:30
Chairperson: Kai Wörner

A linguistics-based speech corpus

Janne Bondi Johannessen, Øystein Alexander Vangsnes, Joel Priestley, Kristin Hagen

In this paper we focus on the linguistic basis for the choices made in the Nordic Dialect Corpus. We focus on transcriptions, annotations, selection of informants, recording situation, user-friendly search interface, links to audio and video, various viewings of results, including maps.

GeWiss – a comparable corpus of academic German, English and Polish

Adriana Slavcheva, Cordula Meißner

The corpus resources available for research on German academic language are limited, even with regard to the written modality. For spoken academic language they are practically non-existent. To make a first step towards remedying this situation, with GeWiss a comparable corpus is being constructed, consisting of spoken academic language data from German, English, and Polish academic contexts. In total it comprises about 120 hours of recording of 447 speakers including native speaker data from Polish, English, and German academics and students, as well as German as a Foreign Language (GFL) data of non-native speakers of German. Data were gathered in two genres (academic papers / student presentations and oral examinations) within one discipline (philology). The GeWiss corpus contains detailed metadata which are stored and administered using the EXMARaLDA Corpus Manager (cf. Schmidt/Wörner 2009). The recordings were transcribed using the EXMARaLDA Partitur Editor (ibid.) and the minimal transcription level of the GAT2 transcription conventions (cf. Selting et al. 2009), which were adapted for the multilingual GeWiss data. Besides the design of the GeWiss corpus, the metadata description and the transcription conventions applied, we describe the workflow from data gathering to corpus publication, which is planned by the end of this year.

Multimodal Parallel Russian Corpus (MultiPARC): Main Tasks and General Structure

Elena Grishina, Svetlana Savchuk, Dmitry Sichinava

The paper introduces a new project, the Multimodal Parallel Russian Corpus, which is planned to be created in the framework of the Russian National Corpus and to include different realizations of the same text: the screen versions and theatrical performances of the same drama, recitations of the same poetical text, and so on. The paper outlines some ways to use the MultiPARC data in linguistic studies.

Constructing General and Dialectal Corpora for Language Variation Research: Two Case Studies from Turkish

Sukriye Ruhi, E. Eda Isik Tas

Parallel corpora utilizing a standardized system for sampling, recording, transcription and annotation potentially ensures cross-linguistic analyses for researchers. This paper describes how comparability was achieved in a general corpus and a dialectal corpus in the context of the Spoken Turkish Corpus – STC (Turkish spoken in Turkey) and the Spoken Turkish Cypriot Dialect Corpus

– STCDC (Turkish in Northern Cyprus). Section 2 overviews aspects of Variety corpora features that impact variation research. Section 3 describes the corpus design and sampling procedures adopted in STC and STCDC. Section 4 focuses on the corpus construction tools employed in the two corpora. Finally, Section 5 presents the transcription standardization methods devised for STC and STCDC.

The Corpus of Spoken Greek: goals, challenges, perspectives

Theodossia-Soula Pavlidou

The purpose of the present paper is to introduce the Corpus of Spoken Greek, which has been developed at the Institute of Modern Greek Studies (Manolis Triandaphyllidis Foundation), Aristotle University of Thessaloniki. More specifically, I would like to describe and account for the particularities of this corpus, which is mainly intended for qualitative research purposes from the perspective of Conversation Analysis. I will thus exemplify some of the issues and challenges involved in the development of corpora that consist of naturalistic speech data. As a consequence, I would like to conclude that the idea of "best practices" for speech corpora can only be understood as a function of the research goals, explicit or implicit, that are prominent when compiling a corpus. Moreover, standardization of speech corpora can only be pursued to an extent that allows, on the one hand, comparability with other corpora and usability by a large community of researchers, and, on the other, ensures maintenance of those characteristics that are indispensable for the kind of research that the corpus was originally conceived for.

Annotating spoken language

Ines Rehbein, Sören Schalowski, Heike Wiese

The design of a linguistically annotated corpus of spoken language is crucial for the future usefulness of the resource, and should thus be carefully tailored towards the needs of the corpus users and the characteristics of the data while, at the same time, paying attention to existing standards to support comparability and interoperability of language resources. In this paper, we outline important issues for the design of a syntactically annotated corpus of spoken language, focussing on standardisation/interoperability and segmentation, and put our proposals up for discussion.

Turn-taking in interdisciplinary scientific research meetings: Using 'R' for a simple and flexible tagging system

Seongsook Choi, Keith Richards

This paper presents our initial step towards identifying and mapping functions (of utterances/turns) and actions (a series of connected actions managed over the course of a sequence of turns) inherent in authentic spoken language data using a simple and flexible tagging system in R. Our ultimate goal is to capture the patterns of dynamic practices through which interactants produce and understand talk-in-interaction both qualitatively and quantitatively. The procedure involves annotating the transcripts with tags that blends elements of CA (conversation analysis) and DA (discourse analysis), which we can then analyse quantitatively. The paper addresses the challenge of developing and annotating a CA and DA integrated tagging system and demonstrates graphical representation of a quantitative analysis that can be derived from it.

Panel: Best Practices

Monday 21 May, 17:00-19:00
Chairperson: Thomas Schmidt

Russian Speech Corpora Framework for Linguistic Purposes

Pavel Skrelin, Daniil Kocharov

The paper introduces a comprehensive speech corpora framework for linguistic purposes developed at the Department of Phonetics, Saint Petersburg State University. It was designed especially for phoneticians, providing them access to speech corpora and convenient tools for speech data selection and analysis. The framework consists of three major parts: speech data, linguistic annotation and software tools for processing and automatic annotation of speech data. The framework was designed for the Russian language. The paper presents the underlying ideas of framework development and describes its architecture.

Developing Solutions for Long-Term Archiving of Spoken Language Data at the Institut für Deutsche Sprache

Peter M. Fischer, Andreas Witt

This document presents ongoing work related to spoken language data within a project that aims to establish a common and unified infrastructure for the sustainable provision of linguistic primary research data at the Institut für Deutsche Sprache (IDS). In furtherance of its mission to "document the German language as it is currently used", the project expects to enable the research community to access a broad empirical base of working material via a single platform. While the goal is to eventually cover all linguistically relevant digital resources of the IDS, including lexicographic information systems such as the IDS German Vocabulary Portal, OWID, written language corpora such as the IDS German Reference Corpus, DeReKo, and spoken language corpora such as the IDS German Speech Corpus for Research and Teaching, FOLK, the work presented here predominantly focuses on the latter type of data, i.e. speech corpora. Within this context, the present document pictures the project's contributions to the development of standards and best practice guidelines concerning data storage, process documentation and legal issues for the sustainable preservation and long-term accessibility of primary linguistic research data.

Using a Global Corpus Data Model for Linguistic and Phonetic Research

Christoph Draxler

This paper presents and discusses the global corpus data model inWikiSpeech for linguistic and phonetic data used at the BAS. The data model is implemented using a relational database system. Two case studies illustrate how the database is used. In the first case study, audio recordings performed via the web in Scotland are accessed to carry out formant analyses of Scottish English vowels. In the second case study, the database is used for online perception experiments on regional variation of speech sounds in German. In both cases, the global corpus data model has shown to an effective means for providing data in the required formats for the tools used in the workflow, and to allow the use of the same database for very different types of applications.

Best Practices in the TalkBank Framework

Brian MacWhinney, Leonid Spektor, Franklin Chen, Yvan Rose

TalkBank is an interdisciplinary research project funded by the National Institutes of Health and the National Science Foundation. The goal of the project is to support data sharing and direct, community-wide access to naturalistic recordings and transcripts of spoken communication. TalkBank has developed consistent practices for data sharing, metadata creation, transcription methods, transcription standards, interoperability, automatic annotation, and dissemination. The database includes corpora from a wide variety of linguistic fields all governed by a comprehensive XML Schema. For each component research subfield, TalkBank must provide special purpose annotations and tools as a subset of the overall system. Together, these various TalkBank standards can serve as guides to further improvements in the use of speech corpora for linguistic research.

Toward the Harmonization of Metadata Practice for Spoken Languages Resources

Christopher Cieri, Malcah Yaeger-Dror

This paper addresses issues related to the elicitation and encoding of demographic, situational and attitudinal metadata for sociolinguistic research with an eye toward standardization to facilitate data sharing. The discussion results from a series of workshops that have recently taken place at the NWAV and LSA conferences. These discussions have focused principally on the granularity of the metadata and the subset of categories that could be considered required for sociolinguistic fieldwork generally. Although a great deal of research on quantitative sociolinguists has taken place in the Unites Stated, the workshops participants actually represent research conducted in North and South America, Europe, Asian, the Middle East, Africa and Oceania. Although the paper does not attempt to consider the metadata necessary to characterize every possible speaker population, we present evidence that the methodological issues and findings apply generally to speech collections concerned with the demographics and attitudes or the speaker pools and the situations under which speech is elicited.

Best Practices in the Design, Creation and Dissemination of Speech Corpora at The Language Archive

Sebastian Drude, Daan Broeder, Peter Wittenburg, Han Sloetjes

In the last 15 years, the Technical Group (now: “The Language Archive”, TLA) at the Max Planck Institute for Psycholinguistics (MPI) has been engaged in building corpora of natural speech and making them available for further research. The MPI has set standards with respect to archiving such resources, and has developed tools that are now widely used, or serve as a reference for good practice. We cover here core aspects of corpus design, annotation, metadata and data dissemination of the corpora hosted at TLA.

Multimodal Corpora:
How Should Multimodal Corpora Deal with the Situation?

22 May 2012

ABSTRACTS

Editors:

Jens Edlund, Dirk Heylen, Patrizia Paggio

Workshop Programme

09:15 – 09:30 – Welcome

09:30 – 10:30 – Session 1

Stylianos Asteriadis, Noor Shaker, Kostas Karpouzis and Georgios N. Yannakakis: *Towards player's affective and behavioral visual cues as drives to game adaptation*

Elena Grishina and Svetlana Savchuk: *Multimodal clusters in spoken Russian*

10:30 – 11:00 Coffee break

11:00 – 12:30 – Session 2

Mary Swift, George Ferguson, Lucian Galescu, Yi Chu, Craig Harman, Hyuckchul Jung, Ian Perera, Young Chol Song, James Allen and Henry Kautz: *A multimodal corpus for integrated language and action*

Masashi Inoue, Ryoko Hanada, Nobuhiro Furuyama, Toshio Irino, Takako Ichinomiya and Hiroyasu Massaki: *Multimodal corpus for psychotherapeutic situation*

Samer Al Moubayed, Jonas Beskow, Björn Granström, Joakim Gustafson, Nicole Mirning, Gabriel Skantze and Manfred Tscheligi: *Furhat goes to Robotville: A large-scale human-robot interaction data collection in a public space*

12:30 – 14:00 Lunch break

14:00 – 16:00 Session 3

Anders Grove: *Automatic analysis of hand movement phases in video speech*

Yasuharu Den and Tomoko Kowaki: *Annotation and preliminary analysis of eating activity in multi-party table talk*

Patrizia Paggio and Costanza Navarretta: *Classifying the feedback function of head movements and face expressions*

Jens Edlund, Mattias Heldner and Joakim Gustafson: *Who am I speaking at? Perceiving the head orientation of speakers from acoustic cues alone*

16:00 – 16:30 Coffee break

16:30 – 17:30 Session 4

Jens Allwood and Elisabeth Ahlsén: *Incremental collection of activity-based multimodal corpora and their use in activity-based studies*

Johannes Wienke, David Klotz and Sebastian Wrede: *A framework for the acquisition of multimodal human-robot interaction data sets with a whole-system perspective*

17:30 – 18:00 Discussion and closing

Workshop Organizers

Jens Edlund	KTH, Sweden
Dirk Heylen	Univ. of Twente, The Netherlands
Patrizia Paggio	Univ. of Copenhagen, Denmark/ Univ. of Malta, Malta

Workshop Programme Committee

Hamid Aghajan	Stanford University, USA
Elisabeth Ahlsen	Univ. of Göteborg, Sweden
Jan Alexandersson	DFKI, Germany
Jens Alwood	Univ. of Göteborg, Sweden
Philippe Blache	Univ. de Provence, France
Susanne Burger	Carnegie Mellon Univ., USA
Kristiina Jokinen	Univ. of Helsinki, Finland
Stefan Kopp	Univ. Bielefeld, Germany
Costanza Navaretta	Univ. Of Copenhagen, Denmark
Karola Pitsch	Univ. Bielefeld, Germany
Andrei Popescu-Belis	Idiap Research Inst., Switzerland
Ronald Poppe	Univ. of Twente, The Netherlands
Albert Ali	Salah Bogazici University, Turkey
Bjoern Schuller	TU Munich, Germany
Alessandro Vinciarelli	Univ. Glasgow, UK

Introduction

Currently, the creation of a multimodal corpus involves the recording, annotation and analysis of a selection of many possible communication modalities such as speech, hand gesture, facial expression, and body posture. Simultaneously, an increasing number of research areas are transgressing from focused single modality research to full-fledged multimodality research. Multimodal corpora are becoming a core research asset and they provide an opportunity for interdisciplinary exchange of ideas, concepts and data.

The 8th Workshop on Multimodal Corpora is again collocated with LREC, which has selected *Speech and Multimodal Resources* as its special topic. This points to the significance of the workshop's general scope, and the fact that the main conference special topic largely covers the broad scope of our workshop provides us with a unique opportunity to step outside the boundaries and look further into the future, and emphasize the fact that a growing segment of research takes a view of spoken language as situated action, where linguistic and non-linguistic actions are intertwined with the dynamic conditions given by the situation and the place in which the actions occur. As a result, the 2012 Workshop on Multimodal Corpora holds a number contributions which share a focus on the acquisition, description, and analysis of situated multimodal corpora.

Welcome

Tuesday 22 May / 9:15 – 9:30

Session 1

Tuesday 22 May / 9:30 – 10:30

Towards player's affective and behavioral visual cues as drives to game adaptation

Stylianos Asteriadis, Noor Shaker, Kostas Karpouzis and Georgios N. Yannakakis

Recent advances in emotion and affect recognition can play a crucial role in game technology. Moving from the typical game controls to controls generated from free gestures is already in the market. Higher level controls, however, can also be motivated by player's affective and cognitive behavior itself, during gameplay. In this paper, we explore player's behavior, as captured by computer vision techniques, and player's details regarding his own experience and profile. The objective of the current research is game adaptation aiming at maximizing player enjoyment. To this aim, the ability to infer player engagement and frustration, along with the degree of challenge imposed by the game is explored. The estimated levels of the induced metrics can feed an engine's artificial intelligence, allowing for game adaptation.

Multimodal clusters in spoken Russian

Elena Grishina and Svetlana Savchuk

The paper introduces the notion of multimodal cluster (MMC). MMC is a multicomponent spoken unit, which includes diads “meaning + gesture”, “meaning + phonetic phenomenon” (double MMC) or triad “meaning + gesture + phonetic phenomenon” (triple MMC). All components of the same MMC are synchronized in the speech, gestural and phonetic components conveying the same idea as the semantic component (naturally, with available means). To put it another way, MMC is a combination of speech phenomena of different modi (semantic, visual, sound), which are tightly connected in the spoken language, and roughly speaking mean the same, i.e. convey the same idea by their own means. The paper describes some examples of double and triple MMCs, which are specific for the Spoken Russian.

Session 2

Tuesday 22 May / 11:00 – 12:30

A multimodal corpus for integrated language and action

Mary Swift, George Ferguson, Lucian Galescu, Yi Chu, Craig Harman, Hyuckchul Jung, Ian Perera, Young Chol Song, James Allen and Henry Kautz

We describe a corpus for research on learning everyday tasks in natural environments using the combination of natural language description and rich sensor data. We have collected audio, video, Kinect RGB-Depth video and RFID object-touch data while participants demonstrate how to make a cup of tea. The raw data are augmented with gold-standard annotations for the language representation and the actions performed. We augment activity observation with natural language instruction to assist in task learning.

Multimodal corpus for psychotherapeutic situation

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

This paper presents a design principle for construction of an in-house multimodal corpus for computationally analysing and better understanding conversations during psychotherapy. We discuss some sharable information about research community data collection procedures such as recording devices, the consent form, and privacy consideration. Also, multimodal coding schema and metadata that are needed in the domain are explained. The corpus has three distinguishing properties: 1) It was constructed only for our own researches and not for public use; 2) The conversation and recording environment was in actual social situations, not controlled; 3) A multimodal coding schema that focuses on the co-construction nature of the conversation was used. Although the conversation contents are not sharable, the data collection procedure and the schema design for the psychotherapy corpus would serve as an example of an initiative to construct a multimodal corpus.

Furhat goes to Robotville: A large-scale human-robot interaction data collection in a public space

Samer Al Moubayed, Jonas Beskow, Björn Granström, Joakim Gustafson, Nicole Mirning, Gabriel Skantze and Manfred Tscheligi

In the four days of the Robotville exhibition at the London Science Museum, UK, during which the back-projected talking head Furhat running a simple yet effective situated spoken dialogue system was seen by almost 8 000 visitors, we collected a database of 16 000 utterances spoken to Furhat in situated and unrehearsed interaction. The data collection is an example of a particular kind of corpus collection of human-machine dialogues in public spaces that has several interesting and specific characteristics, both with respect to the technical details of the collection and with respect to the resulting corpus contents. In this paper, we take the Furhat data collection as a starting point for a discussion of the motives for this type of data collection, its technical peculiarities and prerequisites, and the characteristics of the resulting corpus.

Session 3

Tuesday 22 May / 14:00 – 16:00

Automatic analysis of hand movement phases in video speech

Anders Grove

I find that many video recordings of speeches show rather uncomplicated hand movements and shapes, and regarded as corpus they would fit for analysis by primitive automation. I have implemented an automatic annotation of the hand movement phases and applied it to a political speech on a video clip. The skin colour is used to track the hands, and the boundaries of the phases are settled through changes in speed. For comparison a manual annotation has been made and a set of guidelines stated to ensure the quality and make transparency for evaluation. They are close to the prevailing concept of phase annotation as e.g. stated in the NOVACO scheme (Kipp, 2004), but they also use the hand shape to distinguish the more expressive of the phases. While the automatic annotation is simple, the comparison shows that it is plausible and could be used with caution; the kappa index is a bit above 0.5. A substantial part of the problems origins from the difficulties to distinguish between the hands when they overlap on the screen. If parameters reflecting the form of the hand could be applied it would likely remedy this, and they could also be used in an implementation of the part of the guidelines distinguishing expressive phases based upon the hand forms.

Annotation and preliminary analysis of eating activity in multi-party table talk

Yasuharu Den and Tomoko Kowaki

In this paper, we develop a scheme for annotating eating activity in multi-party table talk, and conduct an initial investigation of how participants coordinate eating and speaking in table talk. In the proposed scheme, eating actions are classified into four categories, i) picking up chopsticks, ii) holding chopsticks, iii) catching food with chopsticks, and iv) taking in food. Each action, then, is sub-divided into a sequence of phases, i.e., preparation, stroke, retraction, and hold, in a similar way as Kendon's gesture annotation scheme. We annotated three 10-minute excerpts from a three-party table-talk corpus in Japanese, and examined the relationship between the time devoted to each type of eating action and participant's engagement in speech activity. Preliminary results showed i) that active speakers tended to spend more time for the ``taking-in-food" action even when they were speaking, and ii) that the hold phase occupied the majority of the time in these ``taking-in-food" actions while speaking. These results suggest that, instead of compensating for the lack of time for eating when they were not speaking, active speakers locally coordinated their eating actions with the speech by halting the movement, and retaining the location, of the hand before putting food in the mouth.

Classifying the feedback function of head movements and face expressions

Patrizia Paggio and Costanza Navarretta

This paper deals with the automatic classification of the feedback function of head movements and facial expressions in the Danish NOMCO corpus, a collection of dyadic conversations in which speakers meet for the first time and speak freely. Two classification tasks are carried out with good results. In the first one, head gestures with a feedback function are learnt. In the second one, the direction of the feedback – whether given or elicited – is predicted. In both cases, we achieve good accuracy (an F-score of 0.764 in the first task and 0.902 in the second), and the best results are obtained when features concerning the shape of both gesture types as well as the words they co-occur with are taken into consideration.

Who am I speaking at? Perceiving the head orientation of speakers from acoustic cues alone

Jens Edlund, Mattias Heldner and Joakim Gustafson

The ability of people, and of machines, to determine the position of a sound source in a room is well studied. The related ability to determine the orientation of a directed sound source, on the other hand, is not, but the few studies there are show people to be surprisingly skilled at it. This has bearing for studies of face-to-face interaction and of embodied spoken dialogue systems, as sound source orientation of a speaker is connected to the head pose of the speaker, which is meaningful in a number of ways. We describe in passing some preliminary findings that led us onto this line of investigation, and in detail a study in which we extend an experiment design intended to measure perception gaze direction to test instead for perception of sound source orientation. The results corroborate those of previous studies, and further show that people are very good at performing this skill outside of studio conditions as well.

Session 4

Tuesday 22 May / 16:30 – 17:30

Incremental collection of activity-based multimodal corpora and their use in activity-based studies

Jens Allwood and Elisabeth Ahlsén

Activity-based communication Analysis is a framework, which puts social activity in focus and analyzes communication in relation to the determining and determined factors of the activity. Given an activity-based approach, it is essential to collect multimodal corpora with a variation of social activities, in order to study similarities, as well as differences between activities and possible influencing factors. The Gothenburg Spoken Language Corpus was collected as a corpus representing communication in a wide range of social activities. The paper describes and briefly discusses the purpose and some of the features of the corpus. The usefulness of activity-based multimodal corpora is exemplified by the analysis of spoken feedback in a specific activity (the physical examination in doctor-patient interaction).

A framework for the acquisition of multimodal human-robot interaction data sets with a whole-system perspective

Johannes Wienke, David Klotz and Sebastian Wrede

In this work we present a conceptual framework for the creation of multimodal data sets which combine human-robot interaction with system-level data from the robot platform. The framework is based on the assumption that perception, interaction modeling and system integration need to be treated jointly in order to improve human-robot interaction capabilities of current robots. To demonstrate the feasibility of the framework, we describe how it has been realized for the recording of a data set with the humanoid robot NAO.

Discussion and closing

Tuesday 22 May / 17:30 – 18:00

Challenges in the Management of Large Corpora

22 May 2012

ABSTRACTS

Editors:

Piotr Bański, Marc Kupietz, Andreas Witt, Damir Ćavar,

Serge Heiden, Anthony Aristar, Helen Aristar-Dry

Workshop Programme

14:00 – Opening

14.03 – 14.30 Keynote talk: Nancy Ide, *Big, Clean, and Comprehensive – but is it Worth it?*

14.30 – 15.00 Lars Bungum and Björn Gambäck, *Efficient N-gram Language Modeling for Billion Word Web-Corpora*

15.00 – 15.30 Hans Martin Lehmann and Gerold Schneider, *Dependency Bank*

15.30 – 16.00 Roman Schneider, *Evaluating DBMS-based access strategies to very large multi-layer corpora*

16:00 – 16:30 Coffee break

16.30 – 17.00 Hanno Biber and Evelyn Breiteneder, *The AAC Container. Managing Text Resources for Text Studies*

17.00 – 17.30 Damir Ćavar, Helen Aristar-Dry and Anthony Aristar, *Large Mailing List Corpora: Management, Annotation and Repository*

17.30 – 18.00 Ritesh Kumar, Pinkey Nainwani, Girish Nath Jha and Shiv Bhusan Kaushik, *Creating and managing large annotated parallel corpora of Indian languages*

18.00 – 18.30 Nelleke Oostdijk and Henk van den Heuvel, *Introducing the CLARIN-NL Data Curation Service*

18.30 – 19.00 Final discussion

Workshop Organizers

Anthony Aristar	Institute for Language Information and Technology, Eastern Michigan University
Helen Aristar-Dry	Institute for Language Information and Technology, Eastern Michigan University
Piotr Bański	Institut für Deutsche Sprache, Mannheim
Damir Ćavar	Institute for Language Information and Technology, Eastern Michigan University
Serge Heiden	ICAR Laboratory, Lyon University
Marc Kupietz	Institut für Deutsche Sprache, Mannheim
Andreas Witt	Institut für Deutsche Sprache, Mannheim

Workshop Programme Committee

Núria Bel	Universitat Pompeu Fabra
Mark Davies	Brigham Young University
Stefanie Dipper	Ruhr-Universität Bochum
Tomaž Erjavec	Jožef Stefan Institute
Stefan Evert	Technische Universität Darmstadt
Alexander Geyken	Berlin-Brandenburgische Akademie der Wissenschaften
Andrew Hardie	University of Lancaster
Nancy Ide	Vassar College
Sandra Kübler	Indiana University
Martin Mueller	Northwestern University
Mark Olsen	University of Chicago
Adam Przepiórkowski	Polish Academy of Sciences, University of Warsaw
Reinhard Rapp	Johannes Gutenberg-Universität Mainz, University of Leeds
Laurent Romary	INRIA, Humboldt-Universität zu Berlin
Pavel Straňák	Charles University in Prague
Amir Zeldes	Humboldt-Universität zu Berlin

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 the 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 or to find and represent fine-grained sub-regularities; on the other hand, some fundamental technical methods and strategies are being called into question. These include e.g. successful curation of the 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.

The workshop on “Challenges in the management of large corpora” aims at gathering the leading researchers in the field of Language Resource creation and Corpus Linguistics, in order to provide for an intensive exchange of expertise, results and ideas.

Keynote: Big, Clean, and Comprehensive – But is it Worth It?

Nancy Ide

Several projects have devoted considerable time, effort, and funding to the development of language corpora, in order to provide large amounts of linguistically annotated data to support natural language processing research and development, and in particular for developing statistical language models that can enable machine learning. As opposed to data collected from the web, these corpora are ”clean” (In the sense of having been rendered into a tractable format for processing), can be enhanced with multiple layers of linguistic annotation, can be designed to cover a “representative” set of genres, and, perhaps most importantly, can be re-distributed for reuse by others–clear advantages that on the face of it seem to justify the effort of constructing these corpora. However, corpus construction is only a first step; to be of real use, the data and annotations must be searchable and accessible via methods that go well beyond simple ”Google search”, thus potentially demanding software development by institutions with limited personnel and funding. Beyond this is the effort required to maintain the corpus and the software and provide for their access and distribution, which in itself can demand a major investment of time and resources. We can even consider efforts to develop standards that might contribute to data and software reuse as another significant cost of large corpus development.

This talk will attempt to weigh the benefits that these resources provide to the natural language processing and linguistics communities against the time, effort, and expense of language resource development, in order to determine whether or not the benefits justify the costs. I will look at the uses to which language corpora are put by these communities and consider the degree to which carefully-constructed, annotated language corpora enable research and development that is quantifiably beyond what could be done using web resources–either existing resources or what we can assume is in the foreseeable future. I will also consider the likelihood that, given their far superior resources, enterprises such as Google and/or projects such as the Semantic Web will eventually render large corpus construction and maintenance unnecessary.

The AAC Container. Managing Text Resources for Text Studies

Hanno Biber and Evelyn Breiteneder

The aim of this paper about the concept of the "AAC container" is to contribute to the workshop theme of managing large corpora by putting emphasis on the perspective of how to come to terms with the actual content of a text corpus by applying approaches based upon the methodologies of text studies. The "AAC-Austrian Academy Corpus" is a large digital text corpus operated by the "Institute for Corpus Linguistics and Text Technology" of the "Austrian Academy of Sciences" in Vienna. Thousands of German language documents and literary objects by thousands of authors have been collected. The historical period covered by this text corpus of 500 million tokens is ranging from the 1848 revolution to the fall of the iron curtain in 1989. In this period significant historical changes with remarkable influences on the language and the language use can be observed. Among the AAC's sources, which cover many domains and genres, there are literary journals, newspapers, novels, dramas, poems, advertisements, essays, travel accounts, cookbooks, pamphlets, political speeches, scientific, legal, religious texts, etc. The AAC corpus holdings provide a great number of reliable resources and interesting corpus based approaches for investigations into the linguistic and textual properties of these texts.

Efficient N-gram Language Modeling for Billion Word Web-Corpora

Lars Bungum and Björn Gambäck

Building higher-order n-gram models over 10s of GB of data poses challenges in terms of speed and memory; parallelization and processing efficiency are necessary prerequisites to build models in feasible time. The paper describes the methodology developed to carry out this task on web-induced corpora within a project aiming to develop a hybrid statistical MT system. Using this parallel processing methodology, a 5-gram LM with Kneser-Ney smoothing for a 3Bn word corpus can be built in half a day. About half of that time is spent in the parallelized part of the process. For a serial execution of the script, this time usage would have had to have been multiplied by 250 (corresponding to close to two months of work).

Large Mailing List Corpora: Management, Annotation and Repository

Damir Ćavar, Helen Aristar-Dry and Anthony Aristar

We present the processes of corpus maintenance, linguistic analysis and annotation, and storage and retrieval infrastructure that is used for the LINGUIST List Mailing List corpus. On the one hand, we describe the setup of text and language processing tools for automatic linguistic annotation, based on common linguistic analysis components as used with GATE or UIMA. On the other hand, we describe the issues and performance evaluations related to storages in relational databases, XML-databases and NoSQL storages.

Creating and managing large annotated parallel corpora of Indian languages

Ritesh Kumar, Shiv Bhusan Kaushik, Pinkey Nainwani and Girish Nath Jha

This paper presents the challenges in creating and managing large parallel corpora of 12 major Indian languages (which is soon to be extended to 23 languages) as part of a major consortium project funded by the Department of Information Technology (DIT), Govt. of India, and running parallel in 10 different universities of India. In order to efficiently manage the process of creation and dissemination of these huge corpora, the web-based (with a reduced stand-alone version also) annotation tool ILCIANN (Indian Languages Corpora Initiative Annotation Tool) has been developed. It was primarily developed for the POS annotation as well as the management of the corpus annotation by people with differing amount of competence and at locations physically situated far apart. In order to maintain consistency and standards in the creation of the corpora, it was necessary that everyone works on a common platform which was provided by this tool.

Dependency Bank

Hans Martin Lehmann and Gerold Schneider

In this paper we present a dependency bank framework that scales from small sets like the ICE corpora to data sets of more than 1000 million words. The dependency bank encodes information at the levels of word-class, chunking and dependency syntax. We discuss the structure of the database, the annotation chain and present a web-based interface. We then discuss potential applications as well as limitations of our fully automatic annotation strategy.

Introducing the CLARIN-NL Data Curation Service

Nelleke Oostdijk and Henk van den Heuvel

In this paper we introduce the CLARIN-NL Data Curation Service. We highlight its tasks and its mediating position between researchers and the CLARIN Data Centres. We outline a scenario for successful data curation and stress the need to take notice of the factors that determine the desirability and feasibility of data curation. Finally, we present and discuss an exemplary case that illustrates the relevant issues involved in setting up a data curation plan.

Evaluating DBMS-based Access Strategies to Very Large Multi-layer Corpora

Roman Schneider

Linguistic query systems are special purpose IR applications. As text sizes, annotation layers, and metadata schemes of language corpora grow rapidly, performing complex searches becomes a highly computational expensive task. We evaluate several storage models and indexing variants in two multi-processor/multi-core environments, focusing on prototypical linguistic querying scenarios. Our aim is to reveal modeling and querying tendencies – rather than absolute benchmark results – when using a relational database management system (RDBMS) and MapReduce for natural language corpus retrieval. Based on these findings, we are going to improve our approach for the efficient exploitation of very large corpora, combining advantages of state-of-the-art database systems with decomposition/parallelization strategies. Our reference implementation uses the German DeReKo reference corpus with currently more than 4 billion word forms, various multi-layer linguistic annotations, and several types of text-specific metadata. The proposed strategy is language-independent and adaptable to large-scale multilingual corpora.

LREC 2012 Workshop on

Language Resource Merging

22May 2012

ABSTRACTS

Editors:

Núria Bel, Maria Gavrilidou, Monica Monachini, Valeria Quochi, Laura Rimell

Workshop Programme

2.00pm – 2.15pm – Welcome and Introduction by Núria Bel

2.15pm – 3.00pm – Invited talk
Iryna Gurevych, *How to UBY – a Large-Scale Unified Lexical-Semantic Resource*

3.00pm – 5.30pm – Oral Session

3.00pm – 3.30pm
Laura Rimell, Thierry Poibeau and Anna Korhonen, *Merging Lexicons for Higher Precision Subcategorization Frame Acquisition*

3.30pm – 4.00pm
Muntsa Padró, Núria Bel and Silvia Necşulescu, *Towards the Fully Automatic Merging of Lexical Resources: A Step Forward*

4.00pm – 4.30pm – Coffee break

4.30pm – 5.00pm
Benoît Sagot and Laurence Danlos, *Merging Syntactic Lexica: The Case for French Verbs*

5.00pm – 5.30pm
Cristina Bosco, Simonetta Montemagni and Maria Simi, *Harmonization and Merging of two Italian Dependency Treebanks*

5.30pm – 5.45pm – Short Break

5.45pm – 6.15pm – Poster Session

> Riccardo Del Gratta, Francesca Frontini, Monica Monachini, Valeria Quochi, Francesco Rubino, Matteo Abrate and Angelica Lo Duca, *L-Leme: An Automatic Lexical Merger Based on the LMF Standard*
>
> Anelia Belogay, Diman Karagiozov, Cristina Vertan, Svetla Koeva, Adam Przepiórkowski, Maciej Ogrodniczuk, Dan Cristea, Eugen Ignat and Polivios Raxis, *Merging Heterogeneous Resources and Tools in a Digital Library*
>
> Thierry Declerck, Stefania Racioppa and Karlheinz Mörth, *Automatized Merging of Italian Lexical Resources*
>
> Radu Simionescu and Dan Cristea, *Towards an Universal Automatic Corpus Format Interpreter Solution*

Workshop Organizers

Núria Bel	Universitat Pompeu Fabra, Barcelona, Spain
Maria Gavrilidou	ILSP/Athena R.C., Athens, Greece
Monica Monachini	CNR-ILC, Pisa, Italy
Valeria Quochi	CNR-ILC, Pisa, Italy
Laura Rimell	University of Cambridge, UK

Workshop Programme Committee

Victoria Arranz	ELDA, Paris, France
Paul Buitelaaar	National University of Ireland, Galway, Ireland
Nicoletta Calzolari	CNR-ILC, Pisa, Italy
Olivier Hamon	ELDA, Paris, France
Aleš Horák	Masaryk University, Brno, Czech Republic
Nancy Ide	Vassar College, Mass. USA
Bernardo Magnini	FBK, Trento, Italy
Paola Monachesi	Utrecht University, Utrecht, The Netherlands
Jan Odijk	Utrecht University, Utrecht, The Netherlands
Muntsa Padró	UPF-IULA, Barcelona, Spain
Karel Pala	Masaryk University, Brno, Czech Republic
Pavel Pecina	Charles University, Prague, Czech Republic.
Thierry Poibeau	University of Cambridge, UK and CNRS, Paris, France
Benoît Sagot	INRIA, Paris, France
Kiril Simov	Bulgarian Academy of Sciences, Sofia, Bulgaria
Claudia Soria	CNR-ILC, Pisa, Italy
Maurizio Tesconi	CNR-IIT, Pisa
Antonio Toral	DCU, Dublin, Ireland

Introduction

The availability of adequate language resources has been a well-known bottleneck for most high-level language technology applications, e.g. Machine Translation, parsing, and Information Extraction, for at least 15 years, and the impact of the bottleneck is becoming all the more apparent with the availability of higher computational power and massive storage, since modern language technologies are capable of using far more resources than the community produces. The present landscape is characterized by the existence of numerous scattered resources, many of which have differing levels of coverage, types of information and granularity. Taken singularly, existing resources do not have sufficient coverage, quality or richness for robust large-scale applications, and yet they contain valuable information (Monachini et al. 2004 and 2006; Soria et al. 2006; Molinero, Sagot and Nicolas 2009; Necşulescu et al. 2011). Differing technology or application requirements, ignorance of the existence of certain resources, and difficulties in accessing and using them, has led to the proliferation of multiple, unconnected resources that, if merged, could constitute a much richer repository of information augmenting either coverage or granularity, or both, and consequently multiplying the number of potential language technology applications. Merging, combining and/or compiling larger resources from existing ones thus appear to be a promising direction to take.

The re-use and merging of existing resources is not altogether unknown. For example, WordNet (Fellbaum, 1998) has been successfully reused in a variety of applications. But this is the exception rather than the rule; in fact, merging, and enhancing existing resources is uncommon, probably because it is by no means a trivial task given the profound differences in formats, formalisms, metadata, and linguistic assumptions.

The language resource landscape is on the brink of a large change, however. With the proliferation of accessible metadata catalogues, and resource repositories (such as the new META-SHARE[1] infrastructure), a potentially large number of existing resources will be more easily located, accessed and downloaded. Also, with the advent of distributed platforms for the automatic production of language resources, such as PANACEA[2], new language resources and linguistic information capable of being integrated into those resources will be produced more easily and at a lower cost. Thus, it is likely that researchers and application developers will seek out resources already available before developing new, costly ones, and will require methods for merging/combining various resources and adapting them to their specific needs.

Up to the present day, most resource merging has been done manually, with only a small number of attempts reported in the literature towards (semi-)automatic merging of resources (Crouch & King 2005; Pustejovsky et al. 2005; Molinero, Sagot and Nicolas 2009; Necsulescu et al. 2011, Gurevych et al. 2012, Eckle-Kohler and Gurevych 2012). In order to take a further step towards the scenario depicted above, in which resource merging and enhancing is a reliable and accessible first step for researchers and application developers, experience and best practices must be shared and discussed, as this will help the whole community avoid any waste of time and resources.

AIMS OF THE WORKSHOP

This half-day workshop is meant to be part of a series of meetings constituting an ongoing forum for sharing and evaluating the results of different methods and systems for the automatic production of language resources (the first one was the LREC 2010 Workshop on Methods for the Automatic Production of Language Resources and their Evaluation Methods). The main focus of this workshop is on (semi-)automatic means of merging language resources, such as lexicons, corpora and grammars. Merging makes it possible to re-use, adapt, and enhance existing resources, alongside new, automatically created ones, with the goal of reducing the manual intervention required in language resource production, and thus ultimately production costs.

[1] http://www.meta-net.eu/meta-share

[2] http://www.panacea-lr.eu/

REFERENCES

Dick Crouch and Tracy H. King. 2005. Unifying lexical resources. *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*. Saarbruecken, Germany.

Judith Eckle-Kohler and Iryna Gurevych. 2012. Subcat-LMF: Fleshing out a standardized format for subcategorization frame interoperability. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, April 2012.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer and Christian Wirth. 2012. Uby - A Large-Scale Unified Lexical-Semantic Resource. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (EACL 2012), April 2012.

Monica Monachini, Nicoletta Calzolari, Khalid Choukri, Jochen Friedrich, Giulio Maltese, Michele Mammini, Jan Odijk & Marisa Ulivieri. 2006. Unified Lexicon and Unified Morphosyntactic Specifications for Written and Spoken Italian. In Calzolari et al. (eds.), *Proceedings of the LREC2006: 5th International Conference on Language Resources and Evaluation*, pp. 1852-1857, Genoa, Italy.

Miguel A. Molinero, Benoît Sagot and Nicolas Lionel. 2009. Building a morphological and syntactic lexicon by merging various linguistic resources. In *Proceedings of 17th Nordic Conference on Computational Linguistics (NODALIDA-09)*, Odense, Danemark

Silvia Necsulescu, Núria Bel, Muntsa Padró, Montserrat Marimon, Eva Revilla. 2011. Towards the Automatic Merging of Language Resources. *First international Workshop on Lexical Resources. Woler 2011*. Ljubljana, Slovenia: 1-5 August 2011.

Pustejovsky, J., M. Palmer and A. Meyers. Towards a Comprehensive Annotation of Linguistic Information. *Workshop on Frontiers in Corpus Annotation II. Pie in the Sky*, ACL, Ann Arbor, MI. 2005.

Claudia Soria, Maurizio Tesconi, Nicoletta Calzolari, Andrea Marchetti, Monica Monachini. 2006. Moving to dynamic computational lexicons with LeXFlow. In *Proceedings of the LREC2006: 5th International Conference on Language Resources and Evaluation*, Genoa, Italy (pp. 7–12).

Oral session

Tuesday 22 May, 3:00pm – 6:00pm
Chairperson: Núria Bel

How to UBY – a Large-Scale Unified Lexical-Semantic Resource

Iryna Gurevych (invited speaker)

The talk will present UBY, a large-scale resource integration project based on the Lexical Markup Framework (LMF, ISO 24613:2008). Currently, nine lexicons in two languages (English and German) have been integrated: WordNet, GermaNet, FrameNet, VerbNet, Wikipedia (DE/EN), Wiktionary (DE/EN), and OmegaWiki. All resources have been mapped to the LMF-based model and imported into an SQL-DB. The UBY-API, a common Java software library, provides access to all data in the database. The nine lexicons are densely interlinked using monolingual and cross-lingual sense alignments. These sense alignments yield enriched sense representations and increased coverage. A sense alignment framework has been developed for automatically aligning any pair of resources mono- or cross-lingually. As an example, the talk will report on the automatic alignment of WordNet and Wiktionary. Further information on UBY and UBY-API is available at: http://www.ukp.tu-darmstadt.de/data/lexical-resources/uby/

Merging Lexicons for Higher Precision Subcategorization Frame Acquisition

Laura Rimell, Thierry Poibeau and Anna Korhonen

We present a new method for increasing the precision of an automatically acquired subcategorization lexicon, by merging two resources produced using different parsers. Although both lexicons on their own have about the same accuracy, using only sentences on which the two parsers agree results in a lexicon with higher precision, without too great loss of recall. This "intersective" resource merger is appropriate when both resources are automatically produced, hence noisy, or when precision is of primary importance, and may also be a useful approach for new domains where sophisticated filtering and smoothing methods are unavailable.

Towards the Fully Automatic Merging of Lexical Resources: a Step Forward

Muntsa Padró, Núria Bel and Silvia Necşulescu

This article reports on the results of our research done towards the fully automatically merging of lexical resources. The main goal is to develop a fully automatic technique that allows the merging of lexical resources in different formats. The proposed technique performs two steps: first of all, it converts both lexica into a common format, automatically learning mapping rules. After, it performs the merging of lexical entries using graph unification. In this work, we aim to show the generality of the proposed approach, which have been previously applied to merge Spanish Subcategorization Frames lexica. Here we extend and apply the same technique to perform the merging of morphosyntactic lexica encoded in LMF. The experiments showed that the technique is general enough to obtain good results in these two different tasks which is an important step towards performing the merging of lexical resources fully automatically.

Merging Syntactic Lexica: the Case for French Verbs

Benoît Sagot and Laurence Danlos

Syntactic lexicons, which associate each lexical entry with information such as valency, are crucial for several natural language processing tasks, such as parsing. However, because they contain rich and complex information, they are very costly to develop. In this paper, we show how syntactic lexical resources can be merged, in order to take benefit from their respective strong points, and despite the disparities in the way they represent syntactic lexical information. We illustrate our methodology with the example of French verbs. We describe four large-coverage syntactic lexicons for this language, among which the Lefff, and show how we were able, using our merging algorithm, to extend and improve the Lefff.

Harmonization and Merging of two Italian Dependency Treebanks

Cristina Bosco, Simonetta Montemagni and Maria Simi

The paper describes the methodology which is currently being defined for the construction of a "Merged Italian Dependency Treebank" (MIDT) starting from already existing resources. In particular, it reports the results of a case study carried out on two available dependency treebanks, i.e. TUT and ISST–TANL. The issues raised during the comparison of the annotation schemes underlying the two treebanks are discussed and investigated with a particular emphasis on the definition of a set of linguistic categories to be used as a "bridge" between the specific schemes. As an encoding format, the CoNLL de facto standard is used.

Poster Session

Tuesday 22 May, 6:00pm – 6.30pm
Chairperson: Núria Bel

L-LEME: an Automatic Lexical Merger based on the LMF standard

Riccardo Del Gratta, Francesca Frontini, Monica Monachini, Valeria Quochi, Francesco Rubino, Matteo Abrate and Angelica Lo Duca

The paper describes L-LEME (LMF LExical MErger), an architecture to combine two lexicons in order to obtain new resource(s). L-LEME relies on standards, thus exploiting the benefits of the ISO Lexical Markup Framework (LMF) to ensure interoperability. L-LEME is meant to be dynamic and heavily adaptable: it allows the users to configure it to meet their specific needs. The L-LEME architecture is composed of two main modules: the Mapper, which takes in input two lexicons, LA and LB and a set of user-defined rules and instructions to guide the mapping process (Directives D), and gives in output all matching entries. The algorithm also calculates a cosine similarity score. The Builder takes in input the previous results, a set of Directives D1 and produces a new LMF lexicon, LC. The Directives allow the user to define its own building rules and different merging scenarios. L-LEME is applied to a specific concrete task within the PANACEA project, namely the merging of two Italian Subcategorization Frame (SCF) lexicons. The experiment is interesting in that LA and LB have different philosophies behind, being A built by human introspection and B automatically extracted. Ultimately, L-LEME has interesting repercussions in many language technology applications.

Merging Heterogeneous Resources and Tools in a Digital Library

Anelia Belogay, Diman Karagiozov, Cristina Vertan, Svetla Koeva, Adam Przepiórkowski, Maciej Ogrodniczuk, Dan Cristea, Eugen Ignat and Polivios Raxis

Merging of Language Resources is not only a matter of mapping between different annotation schemata but also of linguistic tools coping with heterogeneous annotation formats in order to produce one single output. In this paper we present a web content management system ATLAS which succeeded to integrate and harmonize resources and tools for six languages, including four less-resourced ones. As a proof of the concept, we implemented a digital library (i-Librarian). Two user evaluation rounds assessed the users productivity of using a software system harnessing language technologies in the processes of content management.

Automatized Merging of Italian Lexical Resources

Thierry Declerck, Stefania Racioppa and Karlheinz Mörth

In the context of a recently started European project, TrendMiner, there is a need for a large lexical coverage of various languages, among those the Italian language. The lexicon should include morphological, syntactic and semantic information, but also features for representing the level of opinion or sentiment that can be expressed by the lexical entries. Since there is no yet ready to use such lexicon, we investigated the possibility to access and merge various Italian lexical resources. A departure point was the freely available Morph-it! lexicon, which is containing inflected forms with their lemma and morphological features. We transformed the textual format of Morph-it! onto a database schema, in order to support integration process with other resources. We then considered Italian lexicon entries available in various versions of Wiktionary for adding further information, like origin, uses and senses of the entries. We explore the need to have a standardized representation of lexical resources in order to better integrate the various lexical information from the distinct sources, and we also describe a first conversion of the lexical information onto a computational lexicon.

Towards an Universal Automatic Corpus Format Interpreter solution

Radu Simionescu and Dan Cristea

The process of building a processing chain is always cumbersome because, in most cases, the NLP tools making up a chain do not match with respect to the input/output format. Convertors are required to transform the output format of a tool to the input format of the next one in the chain, in order to assure correct communication between modules. The work presented in this paper proposes a solution for automatic format interpretation of annotated corpora. A mechanism of this kind would finally make possible the automatic generation of processing architectures. ALPE is a system designed to compute processing workflows, given a sample of an input format and a description of an intended output format.

Language Technology for Normalisation of Less-Resourced Languages

The 8th International Workshop of the ISCA Special Interest Group
on Speech and Language Technology for Minority Languages (SaLTMiL2012)
and
the 4th Workshop on African Language Technology (AfLaT2012)

22 May 2012

ABSTRACTS

Editors:

Guy De Pauw, Kepa Sarasola and Francis M. Tyers

Workshop Programme

09:15–09:30 Welcome / Opening Session

09:30–10:30 Invited Talk

- Sjur Moshagen Nørstebø. *How to build language technology resources for the next 100 years*

10:30–11:00 Coffee Break

11:00–13:00 Resource Creation

- Elaine Uí Dhonnchadha, Alessio Frenda and Brian Vaughan, *Issues in Designing a Spoken Corpus of Irish.*
- Wondwossen Mulugeta and Michael Gasser, *Learning Morphological Rules for Amharic Verbs Using Inductive LogicProgramming*
- Kristín Bjarnadóttir, *The Database of Modern Icelandic Inflection*
- Fadoua Ataa Allah and Siham Boulaknadel, *Natural Language Processing for Amazigh Language: Challenges and Future Directions*

13:00–14:00 Lunch Break

14:00–16:00 Resource Use

- Tommi A. Pirinen and Francis M. Tyers. *Compiling Apertium morphological dictionaries with HFST and using them in HFST applications.*
- Borbóla Siklósi, György Orosz, Attila Novák and Gábor Prószéky. *Automatic structuring and correction suggestion system for Hungarian clinical records.*
- Linda Wiechetek. *Constraint Grammar based Correction of Grammatical Errors for North Sàmi.*
- Michael Gasser, *Toward a Rule-Based System for English-Amharic Translation.*

16:00–16:30 Coffee Break

16:30–17:30 Poster Session

- Paola Carrión González and Emmanuel Cartier, *Technological Tools for Dictionary and Corpora Building for Minority Languages: Example of the French-based Creoles.*
- Denys Duchier, Brunelle Magnana Ekoukou, Yannick Parmentier, Simon Petitjean and Emannuel Schang, *Describing Morphologically-rich Languages using Metagrammars: a Look at Verbs in Ikota.*
- Tjerk Hagemeijer, Iris Hendrickx, Abigail Tiny and Haldane Amaro, *A Corpus of Santomé.*
- Sigrún Helgadóttir, Asta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir and Hrafn Loftsson, *The Tagged Icelandic Corpus (MM).*
- Laurette Pretorius and Sonja Bosch, Semi-automated extraction of morphological grammars for Nguni with special reference to Southern Ndebele.
- Björn Gambäck, Tagging and Verifying an Amharic News Corpus.
- Guy De Pauw, Gilles-Maurice de Schryver and Janneke van de Loo. *Resource-Light Bantu Part-of-Speech Tagging.*
- Gulshan Dovudov, Vít Suchomel and Pavel Smerk, *POS Annotated 50M Corpus of Tajik Language.*

Workshop Organizers

[Please insert the name(s) and affiliation(s) of the Organizing Committee Members using font Times New Roman, 12 pts]

Guy De Pauw (AfLaT)	CLiPS - Computational Linguistics Group, University of Antwerp, Belgium
Gilles-Maurice de Schryver (AfLaT)	African Languages and Cultures, Ghent University, Belgium Xhosa Department, University of the Western Cape, South Africa
Mikel L. Forcada (SaLTMiL)	Departament de Llenguatges i Sistemes Informàtics, Universitat d'Alacant, Spain
Kepa Sarasola (SaLTMiL)	Dept. of Computer Languages, University of the Basque Country
Francis M. Tyers (SaLTMiL)	Departament de Llenguatges i Sistemes Informàtics, Universitat d’Alacant, Spain
Peter Waiganjo Wagacha (AfLaT)	School of Computing & Informatics, University of Nairobi, Kenya

Workshop Programme Committee

Iñaki Alegria	University of the Basque Country, Spain
Nuria Bel	Universitat Pompeu Fabra, Barcelona, Spain
Lars Borin	Göteborgs universitet, Sweden
Sonja Bosch	University of South Africa, South Africa
Khalid Choukri	ELRA/ELDA, France
Guy De Pauw	Universiteit Antwerpen, Belgium
Gilles-Maurice de Schryver	Universiteit Gent
Mikel L. Forcada	Universitat d’Alacant, Spain
Dafydd Gibbon	Universität Bielefeld, Germany
Lori Levin	Carnegie Mellon University, USA
Hrafn Loftsson	University of Reykjavik, Iceland
Girish Nath Jha	Jawaharlal Nehru University, India
Odétúnjí Odéjobi	Obafemi Awolowo University, Nigeria
Laurette Pretorius	University of South Africa, South Africa
Benoît Sagot	INRIA, France
Felipe Sánchez-Martínez	Universitat d’Alacant, Spain
Kepa Sarasola	University of the Basque Country, Spain
Kevin Scannell	Saint Louis University, United States
Trond Trosterud	University of Tromsø, Norway
Francis M. Tyers	Universitat d’Alacant, Spain
Peter Waiganjo Wagacha	University of Nairobi, Kenya

Preface

The 8th International Workshop of the ISCA Special Interest Group on Speech and Language Technology for Minority Languages (SALTMIL)1 and the Fourth Workshop on African Language Technology (AfLaT2012)2 are jointly held as part of the 2012 International Language Resources and Evaluation Conference (LREC 2012). Entitled “Language technology for normalisation 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) and Malta (2010), and the series of AfLaT workshops, held in Athens (EACL2009), Malta (LREC2010) and Addis Ababa (AGIS11).

The Istanbul 2012 workshop aims to share information on tools and best practices, so that isolated researchers will not need to start their work from scratch. An important aspect will be the forming of personal contacts, which can minimize duplication of effort. There will be a balance between presentations of existing language resources, and more general presentations designed to give background information needed by all researchers.

While less-resourced languages and minority languages often struggle to find their place in a digital world dominated by only a handful of commercially interesting languages, a growing number of researchers are working on alleviating this linguistic digital divide, through localisation efforts, the development of BLARKs (basic language resource kits) and practical applications of human language technologies. The joint SaLTMiL/AfLaT workshop on “Language technology for normalisation of less-resourced languages” provides a unique opportunity to connect these researchers and set up a common forum to meet and share the latest developments in the field.

The workshop takes an inclusive approach to the word normalisation, considering it to include both technologies that help make languages more “normal” in society and everyday life, as well as technologies that normalise languages, i.e. help create or maintain a written standard or support diversity in standards. We particularly focus on the challenges less-resourced and minority languages face in the digital world.

Resource Creation

Tuesday 22 May, 11:00 – 13:00
Chairperson: *Francis M. Tyers*

Issues in Designing a Spoken Corpus of Irish

Elaine Uí Dhonnchadha, Alessio Frenda and Brian Vaughan

Abstract
This paper describes the stages involved in implementing a corpus of spoken Irish. This pilot project (consisting of approximately 140K words of transcribed data) implements part of the design of a larger corpus of spoken Irish which it is hoped will contain approximately 2 million words when complete. It hoped that such a corpus will provide material for linguistic research, lexicography, the teaching of Irish and for development of language technology for the Irish language.

Learning Morphological Rules for Amharic Verbs Using Inductive LogicProgramming

Wondwossen Mulugeta and Michael Gasser

Abstract
This paper presents a supervised machine learning approach to morphological analysis of Amharic verbs. We use Inductive Logic Programming (ILP), implemented in CLOG. CLOG learns rules as a first order predicate decision list. Amharic, an under-resourced African language, has very complex inflectional and derivational verb morphology, with four and five possible prefixes and suffixes respectively. While the affixes are used to show various grammatical features, this paper addresses only subject prefixes and suffixes. The training data used to learn the morphological rules are manually prepared according to the structure of the background predicates used for the learning process. The training resulted in 108 stem extraction and 19 root template extraction rules from the examples provided. After combining the various rules generated, the program has been tested using a test set containing 1,784 Amharic verbs. An accuracy of 86.99% has been achieved, encouraging further application of the method for complex Amharic verbs and other parts of speech.

The Database of Modern Icelandic Inflection

Kristín Bjarnadóttir

Abstract
The topic of this paper is the Database of Modern Icelandic Inflection (DMII), containing about 270,000 paradigms from Modern Icelandic, with over 5.8 million inflectional forms. The DMII was created as a multipurpose resource, for use in language technology, lexicography, and as an online resource for the general public. Icelandic is a morphologically rich language with a complex inflectional system, commonly exhibiting idiosyncratic inflectional variants. In spite of a long history of morphological research, none of the available sources had the necessary information for the making of a comprehensive and productive rule-based system with the coverage needed. Thus, the DMII was created as a database of paradigms showing all and only the inflectional variants of each word. The initial data used for the project was mostly lexicographic. The creation of a 25 million token corpus of Icelandic, the MÍM Corpus, has made it possible to use empirical data in the development of the DMII, resulting in extensive additions to the vocabulary. The data scarcity in the corpus, due to the enormous number of possible inflectional forms, proves how important it is to use both lexicographic data and a corpus to complement each other in an undertaking such as the DMII.

Natural Language Processing for Amazigh Language: Challenges and Future Directions

Fadoua Ataa Allah and Siham Boulaknadel

Abstract
Amazigh language, as one of the indo-European languages, poses many challenges on natural language processing. The writing system, the morphology based on unique word formation process of roots and patterns, and the lack of linguistic corpora make computational approaches to Amazigh language challenging.
In this paper, we give an overview of the current state of the art in Natural Language Processing for Amazigh language in Morocco, and we suggest the development of other technologies needed for the Amazigh language to live in "information society".

Resource Use

Tuesday 22 May, 14:00 – 16:00
Chairperson: *Guy De Pauw*

Compiling Apertium morphological dictionaries with HFST and using them in HFST applications.

Tommi A. Pirinen and Francis M. Tyers

Abstract
In this paper we aim to improve interoperability and re-usability of the morphological dictionaries of Apertium machine translation system by formulating a generic finite-state compilation formula that is implemented in HFST finite-state system to compile Apertium dictionaries into general purpose finite-state automata. We demonstrate the use of the resulting automaton in FST-based spell-checking system.

Automatic structuring and correction suggestion system for Hungarian clinical records.

Borbóla Siklósi, György Orosz, Attila Novák and Gábor Prószéky

Abstract
The first steps of processing clinical documents are structuring and normalization. In this paper we demonstrate how we compensate the lack of any structure in the raw data by transforming simple formatting features automatically to structural units. Then we developed an algorithm to separate running text from tabular and numerical data. Finally we generated correcting suggestions for word forms recognized to be incorrect. Some evaluation results are also provided for using the system as automatically correcting input texts by choosing the best possible suggestion from the generated list. Our method is based on the statistical characteristics of our Hungarian clinical data set and on the HUMor Hungarian morphological analyzer. The conclusions claim that our algorithm is not able to correct all mistakes by itself, but is a very powerful tool to help manually correcting Hungarian medical texts in order to produce a correct text corpus of such a domain.

Constraint Grammar based Correction of Grammatical Errors for North Sámi.

Linda Wiechetek

Abstract
The article describes a grammar checker prototype for North Sámi, a language with agglutinative and inflective features. The grammar checker has been constructed using the rule-based Constraint Grammar formalism. The focus is on the setup of a prototype and diagnosing and correcting grammatical case errors, mostly those that appear with adpositions. Case errors in writing are typical even for native speakers as case errors can result from spelling mistakes. Typical candidates for spelling mistakes are forms containing the letter á and those with double consonants. Alternating double and single consonants is a possible case marker. Case errors in an adpositional phrase are common mistakes. Adpositions are typically homonymous (preposition, postposition, adverb) and ask for a genitive case to the left or right of it. Therefore, finding case errors requires a disambiguation of the adposition itself, a correct dependency mapping between the adposition and its dependent and a diagnosis of the case error, which can require homonymy disambiguation of the dependent itself. A deep linguistic analysis including a module for disambiguation, syntactic analysis and dependency annotation is necessary for correcting case errors in adpositional phrases.

Toward a Rule-Based System for English-Amharic Translation.

Michael Gasser

Abstract
We describe key aspects of an ongoing project to implement a rule-based English-to-Amharic and Amharic-to-English machine translation system within our L3 framework. L3 is based on Extensible Dependency Grammar (Debusmann, 2007), a multi-layered dependency grammar formalism that relies on constraint satisfaction for parsing and generation. In L3 , we extend XDG to multiple languages and translation. This requires a mechanism to handle cross-lingual relationships and mismatches in the number of words between source and target languages. In this paper, we focus on these features as well as the advantages that L3 offers for handling structural divergences between English and Amharic and its capacity to accommodate shallow and deep translation within a single system.

Poster session

Tuesday 22 May, 16:30 – 17:30
Chairperson:

Technological Tools for Dictionary and Corpora Building for Minority Languages: Example of the French-based Creoles.

Paola Carrión González and Emmanuel Cartier

Abstract
In this paper, we present a project which aims at building and maintaining a lexicographical resource of contemporary French-based creoles, still considered as minority languages, especially those situated in American-Caribbean zones. These objectives are achieved through three main steps: 1) Compilation of existing lexicographical resources (lexicons and digitized dictionaries, available on the Internet); 2) Constitution of a corpus in Creole languages with literary, educational and journalistic documents, some of them retrieved automatically with web spiders; 3) Dictionary maintenance: through automatic morphosyntactic analysis of the corpus and determination of the

frequency of unknown words. Those unknown words will help us to improve the database by searching relevant lexical resources that we had not included before. This final task could be done iteratively in order to complete the database and show language variations within the same Creole-speaking community. Practical results of this work will consist in 1/ A lexicographical database, explicitating variations in French-based creoles, as well as helping normalizing the written form of this language; 2/ An annotated corpora that could be used for further linguistic research and NLP applications.

Describing Morphologically-rich Languages using Metagrammars: a Look at Verbs in Ikota.

Denys Duchier, Brunelle Magnana Ekoukou, Yannick Parmentier, Simon Petitjean and Emannuel Schang.

Abstract

In this paper, we show how the concept of metagrammar originally introduced by Candito (1996) to design large Tree-Adjoining Grammars describing the syntax of French and Italian, can be used to describe the morphology of Ikota, a Bantu language spoken in Gabon. Here, we make use of the expressivity of the XMG (eXtensible MetaGrammar) formalism to describe the morphological variations of verbs in Ikota. This XMG specification captures generalizations over these morphological variations. In order to produce the inflected forms, one can compile the XMG specification, and save the resulting electronic lexicon in an XML file, thus favorising its reuse in dedicated applications.

A Corpus of Santomé.

Tjerk Hagemeijer, Iris Hendrickx, Abigail Tiny and Haldane Amaro.

Abstract

We present the process of constructing a corpus of spoken and written material for Santome, a Portuguese-related creole language spoken on the island of S. Tomé in the Gulf of Guinea (Africa). Since the language lacks an official status, we faced the typical difficulties, such as language variation, lack of standard spelling, lack of basic language instruments, and only a limited data set. The corpus comprises data from the second half of the 19 th century until the present. For the corpus compilation we followed corpus linguistics standards and used UTF-8 character encoding and XML to encode meta information. We discuss how we normalized all material to one spelling, how we dealt with cases of language variation, and what type of meta data is used. We also present a POS-tag set developed for the Santome language that will be used to annotate the data with linguistic information.

The Tagged Icelandic Corpus (MM).

Sigrún Helgadóttir, Asta Svavarsdóttir, Eiríkur Rögnvaldsson, Kristín Bjarnadóttir and Hrafn Loftsson,

Abstract

In this paper, we describe the development of a morphosyntactically tagged corpus of Icelandic, the MÍM corpus. The corpus consists of about 25 million tokens of contemporary Icelandic texts collected from varied sources during the years 2006–2010. The corpus is intended for use in Language Technology projects and for linguistic research. We describe briefly other Icelandic corpora and how they differ from the MÍM corpus. We describe the text selection and collection for MÍM, both for written and spoken text, and how metadata was created. Furthermore, copyright issues are discussed and how permission clearance was obtained for texts from different sources. Text cleaning and annotation phases are also described. The corpus is available for search through a web interface and for download in TEI-conformant XML format. Examples are given of the use of the corpus and some spin-offs of the corpus project are described. We believe that the care with which we secured copyright clearance for the texts will make the corpus a valuable resource for Icelandic Language Technology projects. We hope that our work will inspire those wishing to develop similar resources for less-resourced languages.

Semi-automated extraction of morphological grammars for Nguni with special reference to Southern Ndebele.

Laurette Pretorius and Sonja Bosch.

Abstract

A finite-state morphological grammar for Southern Ndebele, a seriously under-resourced language, has been semi-automatically obtained from a general Nguni morphological analyser, which was bootstrapped from a mature hand-written morphological analyser for Zulu. The results for Southern Ndebele morphological analysis, using the Nguni analyser, are surprisingly good, showing that the Nguni languages (Zulu, Xhosa, Swati and Southern Ndebele) display significant cross-linguistic similarities that can be exploited to accelerate documentation, resource-building and software development. The project embraces recognized best practices for the encoding of resources to ensure sustainability, access, and easy adaptability to future formats, lingware packages and development platforms.

Tagging and Verifying an Amharic News Corpus.

Björn Gambäck.

Abstract

The paper describes work on verifying, correcting and retagging a corpus of Amharic news texts. A total of 8715 Amharic news articles had previously been collected from a web site, and part of the corpus (1065 articles; 210,000 words) then morphologically analysed and manually part-of-speech tagged. The tagged corpus has been used as the basis for testing the application to Amharic of machine learning techniques and tools developed for other languages. This process made it possible to spot several errors and inconsistencies in the corpus which has been iteratively refined, cleaned, normalised, split into folds, and partially re-tagged by both automatic and manual means.

Resource-Light Bantu Part-of-Speech Tagging.

Guy De Pauw, Gilles-Maurice de Schryver and Janneke van de Loo.

Abstract

Recent scientific publications on data-driven part-of-speech tagging of Sub-Saharan African languages have reported encouraging accuracy scores, using off-the-shelf tools and often fairly limited amounts of training data. Unfortunately, no research efforts exist that explore which type of linguistic features contribute to accurate part-of-speech tagging for the languages under investigation. This paper describes feature selection experiments with a memory-based tagger, as well as a resource-light alternative approach. Experimental results show that contextual information is often not strictly necessary to achieve a good accuracy for tagging Bantu languages and that decent results can be achieved using a very straightforward unigram approach, based on orthographic features.

POS Annotated 50M Corpus of Tajik Language.

Gulshan Dovudov, Vít Suchomel and Pavel Smerk.

Abstract

Paper presents by far the largest available computer corpus of Tajik language of the size of more than 50 million words. To obtain the texts for the corpus two different approaches were used and the paper offers a description of both of them. Then the paper describes a newly developed morphological analyzer of Tajik and presents some statistics of its application on the corpus.

ColabTKR 2012 - Terminology and Knowledge Representation Workshop

22 May 2012

ABSTRACTS

Editors:

António Lucas Soares, Rute Costa

Workshop Programme

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

14:15 – 14:45

Michael Wetzel, Elena Chiocchetti, Tanja Wissik, *Putting Together Apples and Oranges: The LISE Tool Suite for Collaborative Terminology Work*

14:45 – 15:15

Sérgio Barros, Rute Costa, António Lucas Soares, Manuel Silva, *Integrating terminological methods in a framework for collaborative development of semi-formal ontologies*

15:15 – 15:45

Gabriel Bernier-Colborne, *Defining a Gold Standard for the Evaluation of Term Extractors*

15:45 – 16:15 Coffee break

16:15 – 16:45
Gian Piero Zarri, *Mapping from Lexical Resources to High-Level Data Modelling Languages*

16:45 – 17:15
Cristóvão Sousa, António Lucas Soares, Carla Pereira, Rute Costa, *Supporting the identification of conceptual relations in semi-formal ontology development*

17:15 – 17:30
Conclusions

Workshop Organizers

António Lucas Soares	University of Porto and INESC Porto, Portugal
Rute Costa	New University of Lisbon, CLUNL, Portugal
Carla Pereira	IPP/ESTGF and INESC Porto, Portugal
Alessandro Oltramari	Carnegie-Mellon University, USA
Christophe Roche	University of Savoie, France
Anita Nuopponen	University of Vaasa, Finland

Workshop Programme Committee

Gerhard Budin	University of Vienna
Chiara Ghidini	Bruno Kessler Foundation (FBK) - Trento, Italy
Guadalupe Aguado de Cea	Universidad Politécnica de Madrid
Hanne ErdmanThomsen	Copenhagen Business School
Mustafa Jarrar	University of Birzeit, Palestine
António Lucas Soares	University of Porto and INESC Porto, Portugal
Rute Costa	Universidade Nova de Lisboa, CLUNL, Portugal
Carla Sofia Pereira	Polytechnic Institute of Porto and INESC Porto, Portugal
Alessandro Oltramari	Carnegie-Mellon University, USA
Christophe Roche	University of Savoie, France
Anita Nuopponen	University fo Vaasa, Finland
Piek Vossen	VU University Amsterdam, Netherlands

Preface

Linguistics and ontology studies have a long record of fruitful cooperation. Cross-research in areas such as computational linguistics, natural language processing, information retrieval and ontology development, maintenance and integration have produced a wealth of multidisciplinary theories, methods, models and tools (Roche, 2008) (Staab, 2008) (Costa & Silva, 2008) (Pereira et al. 2009) . More specifically, the relationship between the lexicon (lexical approaches and resources) and ontology development methods and tools, have been recently well explored in research (Huang et al., 2010). On the contrary, the relationship between terminology and ontology studies, in particular in what concerns to the initial phases of ontology development, has not received so much attention from the scientific communities involved.

Furthermore, in diverse professional areas, new challenges are appearing related with information and knowledge management in highly specialised technical domains, under tightly constrained time requirements, unfolding in collaborative networking contexts. Short-term collaborative networking between individuals, groups and organisations, is recognised by researchers and practitioners as possible solution to cope with an increasingly complex social and economic business environment. Moreover, the current demand for continuous innovation leads to an higher heterogeneity in the technical and scientific domains simultaneously involved in collaborative projects and activities (e.g involving SMEs and research centres) (Camarinha-Matos, 2006). Managing information and knowledge in this context, places new and interesting challenges to terminology and knowledge representation, particularly when these challenges are seen from an integrated terminology/knowledge representation perspective.

Terminological or ontological approaches alone are not likely to be enough in answering to the needs of precision and detail of the specialised technical domains, as much as the research efforts of articulated terminology/ontology approaches are likely to be inadequate in terms of the required resources (time and persons). Thus, these challenges call for more than the setup and configuration of common terminological or ontological resources, particularly when considering the usually accepted time-frames for developing semantic and terminological artifacts. Effective ways to collaboratively construct shared conceptualisations by the means of negotiation and representational artifacts, such as semi-formal ontologies, are then required.

The above problems and difficulties motivate challenging multi and transdisciplinary lines of research in particular where terminology and knowledge representation meet together with a double aim: to collaboratively study the phenomena from cross-perspectives and to produce practical artifacts for professional work in these two areas. This was the motivation for creating the colabTKR - Collaboration in Terminology and Knowledge Representation - workshop where terminology, information/knowledge management, ontology development, and collaboration specialists join to debate and share from problematic theoretical issues to proposals for innovative approaches. ColabTKR main subject - the interplay between terminology and knowledge representation methods and techniques in contexts of collaborative work - encompasses research in topics such as collaborative processes in terminology work, collaborative conceptualization processes and representations of knowledge, multimodal corpora for semi-formal ontology development, theory, methods and tools for conceptual negotiation, interfaces between terminology work and ontology development/maintenance.

In this workshop five papers dealing with different approaches to the collaboration within and between terminology and knowledge representation are presented, three of them describing methods and results obtained in two different projects: LISE project (http://www-lise-termservices.eu) and CogniNet (http://cogninet.tk/).

In the first case, as the authors Michael Wetzel, Elena Chiocchetti, Tanja Wissik, explain in their abstract, the LISE project aims at improving the quality of existing terminology collections and at facilitating the consolidation of administrative nomenclatures and legal terminology. To that purpose, tools and best practices are developed to enhance interoperability and cross-border collaboration, thus offering specific tools to assist the terminological workflow and also a platform to discuss and exchange data.

In the second case, a collaborative platform - ConceptME - was developed under the project cogniNET, a project addressing problems raised by information and knowledge sharing in the context of short life-cycle collaborative networks. The tool provides support to domain experts engaged in activities related to a shared conceptualization. Two presentations were held held regarding ConceptME, as part of the research developed by António Lucas Soares, Rute Costa, Carla Pereira, Sérgio Barros, Cristóvão Sousa, Manuel Silva. The first one deals with the support to the identification of conceptual relations during the development on semi-formal ontologies. The second one describes the integration of a terminological method to support experts in eliciting and organizing concepts of their domain.

In another presentation, Gabriel Benier-Colborne describes a methodology to define a gold standard (fully annotated corpus) for the automatic evaluation of term extractors that he considers relevant to evaluate protocol for term extraction systems.

Finally, Gian Piero Zarri presents a modelling and development tool – NKRL - bringing to discussion the theoretical and practical problems of transferring lexical information to ontological and knowledge-based systems.

The organizers hope that the selection of papers presented here will be of interest to a broad audience, and will be a starting point for further discussion and cooperation.

The Editors
António Lucas Soares
Rute Costa

Thursday 22 May / 14:00 – 17:30
Chairs: António Lucas Soares and Rute Costa

Putting Together Apples and Oranges: The LISE Tool Suite for Collaborative Terminology Work

Michael Wetzel, Elena Chiocchetti, Tanja Wissik

Abstract
Different terminology databases contain different types of information or a diverging depth of information. To create more complete resources, it might be useful to add languages to existing collections and/or merge (part of) some terminology repositories. This being a daunting task in terms of time and staff efforts, tools allowing the semi-automatic processing of data when adding languages, cleaning termbanks from multiple entries or harmonising terminology collections would facilitate this task. The LISE project (http://www.lise-termservices.eu) aims at improving the quality of existing terminology collections and at facilitating the consolidation of administrative nomenclatures and legal terminologies. It develops tools and best practices to enhance interoperability and cross border collaboration. The main purpose is to help terminology managers in public institutions or private service providers and companies improve the coherence and completeness of their terminological resources in a more efficient way. LISE offers specific tools to assist the terminology workflow, but also a platform to discuss and exchange data. The scientific basis of the project rests in a deep insight into terminology workflow best practices, so as to understand at what point in time each specific tool might be usefully applied.

Integrating terminological methods in a framework for collaborative development of semi-formal ontologies

Sérgio Barros, Rute Costa, António Lucas Soares, Manuel Silva

Abstract
Despite the availability of tools, resources and techniques aimed at the construction of ontological artifacts, developing a shared conceptualization of a given reality still raises questions about the principles and methods that support the initial phases of conceptualization.

To tackle this issue a collaborative platform was developed where terminological and knowledge representation processes support domain experts throughout a conceptualization framework.

In this article we describe the integration of a terminological method to support experts in eliciting and organizing concepts of their domain.

The method is based on a linguistic analysis of textual resources with the help of a term extraction tool and by highlighting markers of relations between concepts. An application scenario is then presented to illustrate the connection between the terminological processes and the knowledge representation processes without blurring the theoretical distinction between terms and concepts.

Defining a Gold Standard for the Evaluation of Term Extractors

Gabriel Bernier-Colborne

Abstract

We describe a methodology for constructing a gold standard for the automatic evaluation of term extractors, an important step toward establishing a much-needed evaluation protocol for term extraction systems. The gold standard proposed is a fully annotated corpus, constructed in accordance with a specific terminological setting (i.e. the compilation of a specialized dictionary of automotive mechanics), and accounting for the wide variety of realizations of terms in context. A list of all the terminological units in the corpus is extracted, and may be compared to the output of a term extractor, using a set of metrics to assess its performance. Subsets of terminological units may also be extracted, due to the use of XML for annotation purposes, providing a level of customization. Particular attention is paid to the criteria used to select terminological units in the corpus, and the protocol established to account for terminological variation within the corpus.

Mapping from Lexical Resources to High-Level Data Modelling Languages

Gian Piero Zarri

Abstract

This paper deals with some theoretical and practical problems involved in the transfer from an ‘external’, lexical level to a ‘deep’, conceptual one. The main thesis defended in the paper is linked with the remark that the ‘lexical information’ (in the most general meaning of these words) used to feed the ontological and knowledge-based systems after the passage through some sort of knowledge representation system is not homogeneous from a syntactic and semantic point of view. The recourse to a unique conceptual representation model (the well-known ‘uniqueness syndrome’) is then methodologically erroneous. In NKRL (Narrative Knowledge Representation Language), for example, several representation models are used. The usual “binary” model is utilized for the ‘standard’ NKRL ontology, HClass (ontology of classes). An “n-ary” model, based on the notions of “conceptual predicate” and “functional roles” is used for representing the nodes of HTemp (ontology of templates, i.e., the NKRL “ontology of events”). Recursive lists of (reified) symbolic labels are used for modelling the “connectivity phenomena” and for representing correctly full narratives, complex events, multifaceted eChronicles etc.; special representations are employed for representing the temporal phenomena, and so on.

Supporting the identification of conceptual relations in semi-formal ontology development

Cristóvão Sousa, António Lucas Soares, Carla Pereira, Rute Costa

Abstract

Conceptualisation processes are pervasive to most technical and professional activities, but are seldom addressed explicitly due to the lack of theoretical and practical methods and tools. However, it seems not to be a popular research topic in knowledge representation or its sub-areas such as ontology engineering. The approach described in this paper is a contribution to the development of computer based tools supporting collaborative conceptualisation processes. The particularly challenging problem of conceptual relations elicitation is here tackled through a combination of ontological and terminological analysis, through a double theoretical perspective. A conceptual relations reference model was synthesised from a foundational ontological analysis and implemented through conceptual relations templates. The later are part of the conceptME system, a platform developed within this research line, providing knowledge and terminological tools and resources to support activities that involve collaborative conceptualisation processes. The work described in this paper adds more support to an area where this support is very scarce.

LREC'2012 Workshop: LRE-Rel

Language Resources and Evaluation for Religious Texts

Tuesday 22 May 2012

ABSTRACTS

Editors:

Eric Atwell, Claire Brierley, Majdi Sawalha

LRE-Rel Workshop Programme

Tuesday 22 May 2012

09:00 – 10:30 – Session 1 Papers

09:00 Eric Atwell, Claire Brierley, and Majdi Sawalha (Workshop Chairs)
Introduction to Language Resources and Evaluation for Religious Texts

09.10 Harry Erwin and Michael Oakes
Correspondence Analysis of the New Testament

09.30 Mohammad Hossein Elahimanesh, Behrouz Minaei-Bidgoli and Hossein Malekinezhad
Automatic classification of Islamic Jurisprudence Categories

09.50 Nathan Ellis Rasmussen and Deryle Lonsdale
Lexical Correspondences Between the Masoretic Text and the Septuagint

10.10 Hossein Juzi, Ahmed Rabiei Zadeh, Ehsan Baraty and Behrouz Minaei-Bidgoli
A new framework for detecting similar texts in Islamic Hadith Corpora

10:30 – 11:20 Coffee break and Session 2 Posters

Majid Asgari Bidhendi, Behrouz Minaei-Bidgoli and Hosein Jouzi
Extracting person names from ancient Islamic Arabic texts

Assem Chelli, Amar Balla and Taha Zerrouki
Advanced Search in Quran: Classification and Proposition of All Possible Features

Akbar Dastani, Behrouz Minaei-Bidgoli, Mohammad Reza Vafaei and Hossein Juzi
An Introduction to Noor Diacritized Corpus

Karlheinz Mörth, Claudia Resch, Thierry Declerck and Ulrike Czeitschner
Linguistic and Semantic Annotation in Religious Memento Mori Literature

Aida Mustapha, Zulkifli Mohd. Yusoff and Raja Jamilah Raja Yusof
The Qur'an Corpus for Juzuk Amma

Mohsen Shahmohammadi, Toktam Alizadeh, Mohammad Habibzadeh Bijani and Behrouz Minaei
A framework for detecting Holy Quran inside Arabic and Persian texts

Gurpreet Singh
Letter-to-Sound Rules for Gurmukhi Panjabi (Pa): First step towards Text-to-Speech for Gurmukhi

Sanja Stajner and Ruslan Mitkov
Style of Religious Texts in 20th Century

Daniel Stein
Multi-Word Expressions in the Spanish Bhagavad Gita, Extracted with Local Grammars Based on Semantic Classes

Nagwa Younis
Through Lexicographers' Eyes: Does Morphology Count in Making Qur'anic Bilingual Dictionaries?

Taha Zerrouki, Ammar Balla
Reusability of Quranic document using XML

11:20 – 13:00 – Session 3 Papers

11.20 Halim Sayoud
Authorship Classification of two Old Arabic Religious Books Based on a Hierarchical Clustering

11.40 Liviu P. Dinu, Ion Resceanu, Anca Dinu and Alina Resceanu
Some issues on the authorship identification in the Apostles' Epistles

12.00 John Lee, Simon S. M. Wong, Pui Ki Tang and Jonathan Webster
A Greek-Chinese Interlinear of the New Testament Gospels

12.20 Soyara Zaidi, Ahmed Abdelali, Fatiha Sadat and Mohamed-Tayeb Laskri
Hybrid Approach for Extracting Collocations from Arabic Quran Texts

12.40 Eric Atwell, Claire Brierley, and Majdi Sawalha (Workshop Chairs)
Plenary Discussion

13:00 End of Workshop

Workshop Organizers

AbdulMalik Al-Salman:	King Saud University, Saudi Arabia
Eric Atwell:	University of Leeds, UK
Claire Brierley:	University of Leeds, UK
Azzeddine Mazroui:	Mohammed First University, Morocco
Majdi Sawalha:	University of Jordan, Jordan
Abdul-Baquee Muhammad Sharaf:	University of Leeds, UK
Bayan Abu Shawar:	Arab Open University, Jordan

Workshop Programme Committee

Nawal Alhelwal:	Arabic Department, Princess Nora bint Abdulrahman University, Saudi Arabia
Qasem Al-Radaideh:	Computer Information Systems, Yarmouk University, Jordan
AbdulMalik Al-Salman:	Computer and Information Sciences, King Saud University, Saudi Arabia
Eric Atwell:	School of Computing, University of Leeds, UK
Amna Basharat:	Foundation for Advancement of Science and Technology, FAST-NU, Pakistan
James Dickins:	Arabic and Middle Eastern Studies, University of Leeds, UK
Kais Dukes:	School of Computing, University of Leeds, UK
Mahmoud El-Haj:	Computer Science and Electronic Engineering, University of Essex, UK
Nizar Habash:	Center for Computational Learning Systems, Columbia University, USA
Salwa Hamada:	Electronics Research Institute, Egypt
Bassam Hasan Hammo:	Information Systems, King Saud University, Saudi Arabia
Dag Haug:	Philosophy, Classics, History of Art and Ideas, University of Oslo, Norway
Moshe Koppel:	Department of Computer Science, Bar-Ilan University, Israel
Rohana Mahmud:	Computer Science and Information Technology, University of Malaya, Malaysia
Azzeddine Mazroui:	Mathematics and Computer Science, Mohammed 1st University, Morocco
Tony McEnery:	English Language and Linguistics, University of Lancaster, UK
Aida Mustapha:	Computer Science and Information Technology, University of Putra, Malaysia
Mohamadou Nassourou:	Computer Philology and Modern German Literature, University of Würzburg, Germany
Nils Reiter:	Department of Computational Linguistics, Heidelberg University, Germany
Abdul-Baquee M. Sharaf:	School of Computing, University of Leeds, UK
Bayan Abu Shawar:	Information Technology and Computing, Arab Open University, Jordan
Andrew Wilson:	Linguistics and English Language, University of Lancaster, UK
Nagwa Younis:	English Department, Ain Shams University, Egypt
Wajdi Zaghouani:	Linguistic Data Consortium, University of Pennsylvania, USA

Introduction to
Language Resources and Evaluation for Religious Texts

Eric Atwell, Claire Brierley, and Majdi Sawalha (Workshop Chairs)

Welcome to the first LRE-Rel Workshop on Language Resources and Evaluation for Religious Texts, part of the LREC'2012 Language Resources and Evaluation Conference in Istanbul, Turkey. The focus of this workshop is the application of computer-supported and Text Analytics techniques to religious texts ranging from: the faith-defining religious canon; authoritative interpretations and commentary; sermons; liturgy; prayers; poetry; and lyrics. We see this as an inclusive and cross-disciplinary topic, and the workshop aims to bring together researchers with a generic interest in religious texts to raise awareness of different perspectives and practices, and to identify some common themes.

We therefore welcomed submissions on a range of topics, including but not limited to:

- 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;
- measuring semantic relatedness between multiple 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-supported methods for analysing religious texts.

Our own research has focussed on the Quran (e.g. see Proceedings of the main Conference, LREC'2012); but we were pleased to receive papers dealing with a range of other religious texts, including Muslim, Christian, Jewish, Hindu, and Sikh holy books, as well as religious writings from the 17th and 20th centuries. Many of the papers present an analysis technique applied to a specific religious text, which could also be relevant to analysis of other texts; these include text classification, detecting similarities and correspondences between texts, authorship attribution, extracting collocations or multi-word expressions, stylistic analysis, Named Entity recognition, advanced search capabilities for religious texts, developing translations and dictionaries.

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

Tuesday 22 May, 9:00 – 10:30

Correspondence Analysis of the New Testament

Harry Erwin and Michael Oakes

In this paper we describe the multivariate statistical technique of correspondence analysis, and its use in the stylometric analysis of the New Testament. We confirm Mealand's finding that texts from Q are distinct from the remainder of Luke, and find that the first 12 chapters of Acts are more similar to each other than to either Luke or the rest of Acts. We describe initial work in showing that a possible "Signs Gospel", describing Jesus' seven public miracles, is indeed distinct from the remainder of John's gospel, but that the differences are slight and possibly due to differences in genre.

Automatic classification of Islamic Jurisprudence Categories

Mohammad Hossein Elahimanesh, Behrouz Minaei-Bidgoli and Hossein Malekinezhad

This paper evaluates some of text classification methods to classify Islamic jurisprudence classes. One of prominent Islamic sciences is jurisprudence, which explores the religious rules from religious texts. For this study the Islamic Jurisprudence corpus is used. This corpus consists of more than 17000 text documents covering 57 different categories. The major purpose of this paper is evaluating text to numerical vectors converting methods and evaluating different methods of calculating proximity matrix between text documents for religious text classification. The results indicate that the best classification efficacy is achieved especially when 3-grams indexing method and KNN classifier using cosine similarity measure are applied. We reported 87.3% performance for Islamic jurisprudence categories classification.

Lexical Correspondences Between the Masoretic Text and the Septuagint

Nathan Ellis Rasmussen and Deryle Lonsdale

This paper describes a statistical approach to Biblical vocabulary, in which the parallel structure of the Greek Septuagint and the Hebrew Masoretic text is used to locate correspondences between lexical lemmata of the two Biblical languages and score them with a log-likelihood ratio. We discuss metrics used to propose and select possible correspondences, and include an examination of twenty pre-selected items for recall and of twenty items just above the cutoff for precision. We explore the implications for textual correlation and translation equivalence.

A new framework for detecting similar texts in Islamic Hadith Corpora

Hossein Juzi, Ahmed Rabiei Zadeh, Ehsan Baraty and Behrouz Minaei-Bidgoli

Nowadays similarity detection is one of the most applicable aspects of text mining techniques. There are different methods for similarity detection. This paper presented a new system for text similarity detection in Islamic Large Hadith Corpus of Computer Research Center of Islamic Science (CRCIS). This system used N-gram method and Cosine measure for similarity detection. According to evaluation result, computer-based similarity detection systems can be more efficient than previous related work in text similarity detection. We have obtained a 97% F-Score of similarity detection for Hadith texts. We hope that our system enable researches for finding the unified Hadiths as well as detecting of how one large Hadith is divided into several small Hadiths in different traditional Hadith books. This system would be very fruitful for many researches in the area of Hadith and Holy Quran investigations.

Session 2 Posters

Tuesday 22 May, 10:30 – 11:20

Extracting person names from ancient Islamic Arabic texts

Majid Asgari Bidhendi, Behrouz Minaei-Bidgoli and Hosein Jouzi

Recognizing and extracting name entities like person names, location names, date and time from an electronic text is very useful for text mining tasks. Correct named entity recognition is a vital requirement in resolving problems in modern fields like question answering, abstracting systems, information retrieval, information extraction, machine translation, video interpreting and semantic web searches. In recent years many researches in named entity recognition task has been lead to very good results in English and other European languages; whereas the results are not convincing in other languages like Arabic, Persian and many of South Asian languages. One of the most necessary and problematic subtasks of named entity recognition is person name extracting. In this article we have introduced a system for person name extraction in Arabic religious texts using proposed ``Proper Name candidate injection'' concept in a conditional random fields model. Also we have created a corpus from ancient Arabic religious texts. Experiments have shown very hight efficient results are obtained using this method.

Advanced Search in Quran: Classification and Proposition of All Possible Features

Assem Chelli, Amar Balla and Taha Zerrouki

This paper contains a listing for all search features in Quran that we have collected and a classification depending on the nature of each feature. It's the first step to design an information retrieval system that fits to the specific needs of the Quran.

An Introduction to Noor Diacritized Corpus

Akbar Dastani, Behrouz Minaei-Bidgoli, Mohammad Reza Vafaei and Hossein Juzi

This article is aimed to introduce Noor Diacritized Corpus which includes 28 million words extracted from about 360 hadith books. Despite lots of attempts to diacritize the holy Quran, little diacritizing efforts have been done about hadith texts. This corpus is therefore from a great significance. Different statistical aspects of the corpus are explained in this article. This paper states challenges of diacritizing activities in Arabic language in addition to general specifications of the corpus.

Linguistic and Semantic Annotation in Religious Memento Mori Literature

Karlheinz Mörth, Claudia Resch, Thierry Declerck and Ulrike Czeitschner

The project described in this paper was at first concerned with the specific issue of annotating historical texts belonging to the Memento mori genre. To produce a digital version of these texts that could be used to answer the specific questions of the researchers involved, a multi-layered approach was adopted: Semantic annotations were applied to the digital text corpus in order to create a domain-specific taxonomy and thus to facilitate the development of innovative approaches in both literary and linguistic research. In addition, the project aimed to develop text technological methodologies tailored to this particular type of text. This work can be characterised by a high degree of interdisciplinarity, as research has been carried out with both a literary/historical and linguistic/lexicographic perspective. The annotations created as part of this project were also designed to be used in the adaptation of existing linguistic computational tools to suit the needs of non-canonical language varieties.

The Qur’an Corpus for Juzuk Amma

Aida Mustapha, Zulkifli Mohd. Yusoff and Raja Jamilah Raja Yusof

This paper presents a corpus that offers rich knowledge for Juz’ Amma. The corpus is designed to be in dual-language, which are English and Malay. The knowledge covers translation for each word and verse, tafsir, as well as hadith from different authenticated sources. This corpus is designed to support dialogue interaction with an information visualization system for Quranic text called AQILAH. This corpus is hoped to aid mental visualization in studying the Qur’an and to enable the users to communicate the content of Juz’ Amma with clarity, precision, and efficiency.

A framework for detecting Holy Quran inside Arabic and Persian texts

Mohsen Shahmohammadi, Toktam Alizadeh, Mohammad Habibzadeh Bijani and Behrouz Minaei

This paper presents how to design and implement the Quranic intelligent engine to detect Quranic verses in the texts automatically. Process area of this system is in the scope of text mining processes and its operations are beyond the usual multiple patterns matching for reasons are explained in the paper. A new algorithm based on indexing text and patterns is designed in implementation of this system in which the main idea is to map text and patterns to some numerical arrays and process on them rather than the text. This algorithm detects Quranic verses in two stages. In the first stage, using a new index based exact multiple patterns matching algorithm, Quranic words in the text are detected and are converted to numerical arrays. In the second stage a filter is designed that by search on the arrays is able to detect the indexing sequence between arrays and determine whether these words are a part of Quran. Experimental results show that processing on numerical values rather than the text has a significant impact on increasing the performance of algorithm to be faster and more precise for detecting holy Quran phrases inside the texts.

Letter-to-Sound Rules for Gurmukhi Panjabi (Pa): First step towards Text-to-Speech for Gurmukhi

Gurpreet Singh

This article presents the ongoing work to develop Lettert-to-Sound rules for Guru Granth Sahib, the religious scripture of Sikh religion. The corpus forming the basis for development of the rules is taken from EMILLE corpora. Guru Granth Sahib is collection of hyms by founders of Sikh religion. After presenting an overview of Guru Granth Sahib and IPA representation in section 1 and Text-to-Speech in section 2, Letter-to-Sound rules developed will be presented in section 3. This paper will close with final discussion and future directions in section 4. The work presented stand at the development stage and no testing or experiment have so far been performed. The intention is to develop the Text-to-Speech for Punjabi language after developing it for limited set of language available in Guru Granth Sahib.

Style of Religious Texts in 20th Century

Sanja Stajner and Ruslan Mitkov

In this study, we present the results of the investigation of diachronic stylistic changes in 20th century religious texts in two major English language varieties -- British and American. We examined a total of 146 stylistic features, divided into three main feature sets: (average sentence length, Automated readability index, lexical density and lexical richness), part-of-speech frequencies and stop-words frequencies. All features were extracted from the raw text version of the corpora, using the state-of-the-art NLP tools and techniques. The results reported significant changes of various stylistic features belonging to all three aforementioned groups in the case of

British English (1961--1991) and various features from the second and third group in the case of American English (1961--1992). The comparison of diachronic changes between British and American English pointed out very different trends of stylistic changes in these two language varieties. Finally, the applied machine learning classification algorithms indicated the stop-words frequencies as the most important stylistic features for diachronic classification of religious texts in British English and made no preferences between the second and third group of features in diachronic classification in American English.

Multi-Word Expressions in the Spanish Bhagavad Gita, Extracted with Local Grammars Based on Semantic Classes

Daniel Stein

As religious texts are often composed in metaphorical and lyrical language, the role of multi-word expressions (MWE) can be considered even more important than usually for automatic processing. Therefore a method of extracting MWE is needed, that is capable of dealing with this complexity. Because of its ability to model various linguistic phenomena with the help of syntactical and lexical context, the approach of Local Grammars by Maurice Gross seems promising in this regard. For the study described in this paper I evaluate the use of this method on the basis of a Spanish version of the Hindu poem Bhagavad Gita. The search will be refined on nominal MWE, i.e. nominal compounds and frozen expressions with two or three main elements. Furthermore, the analysis is based on a set of semantic classes for abstract nouns, especially on the semantical class “phenomenon”. In this article, the theory and application of Local Grammars is described, and the very first results are discussed in detail.

Through Lexicographers’ Eyes: Does Morphology Count in Making Qur’anic Bilingual Dictionaries?

Nagwa Younis

The existence of several forms of the same word in the Holy Quran is regarded as a source of difficulty for lexicographers who are interested in making Qur’anic bilingual dictionaries. Modern dictionaries nowadays use a considerable amount of corpora to illustrate the actual contexts in which a certain form of a word occurs. The study surveys the attempts of finding equivalents for these forms in the on-line Qur’anic Dictionary provided in the Quranic Corpus (Dukes, 2011). The results of the study shed light on some linguistic aspects in the making of a specialised dictionary of the Holy Quran using a corpus-based approach. These insights are of importance both in the field of lexicography in general and the making of a specialised bilingual Qura’nic dictionary in particular.

Reusability of Quranic document using XML

Taha Zerrouki, Ammar Balla

In this paper, we present Quranic document modelling with XML technologies that offer the benefit of reusability. The use of this technique presents a lot of advantages in Quranic application development such as speed of development, low development cost and being readily available. Existing models fell short of combining these features. The paper reviews the existing models and details the potential of each of the models from perspectives of applications, reusability, and availability.

Session 3 Papers

Tuesday 22 May, 11:20 – 13:00

Authorship Classification of Two Old Arabic Religious Books Based on a Hierarchical Clustering

Halim Sayoud

Authorship classification consists in assigning classes to a set of different texts, where each class should represent one author. In this investigation, the author presents a stylometric research work consisting in an automatic authorship classification of eight different text segments corresponding to four text segments of the Quran (The holy words and statements of God in the Islamic religion) and four other text segments of the Hadith (statements said by the prophet Muhammad). Experiments of authorship attribution are made on these two old religious books by employing a hierarchical clustering and several types of original features. The sizes of the segments are more or less in the same range. The results of this investigation shed light on an old religious enigma, which has not been solved for fifteen hundred years: results show that the two books should have two different authors or at least two different writing styles.

Some issues on the authorship identification in the Apostles' Epistles

Liviu P. Dinu, Ion Resceanu, Anca Dinu and Alina Resceanu

The New Testament is a collection of writings having multiple authors. The traditional view, that of the Christian Church, is that all the books were written by apostles (Mathew, Paul, Peter, John) or by the apostles' disciples (Mark and Luke). However, with the development of literary theory and historical research, this point of view is disputed. For example, the authorship of seven out of the fourteen epistles canonically attributed to Paul is questioned by modern scholars: the Pastoral epistles (First Timothy, Second Timothy, and Titus), which are thought to be pseudoepigraphic, another three about which modern scholars are evenly divided (Ephesians, Colossians, Second Thessalonians), and the anonymous Hebrews, which, most scholars agree, wasn't written by Paul, but may have been written by a disciple of Paul's. In this paper we applied two different techniques (PCA and clustering based on rank distance) to investigate the authorship identification of Apostles' Epistles.

A Greek-Chinese Interlinear of the New Testament Gospels

John Lee, Simon S. M. Wong, Pui Ki Tang and Jonathan Webster

This paper describes an interlinear text consisting of the original Greek text of the four gospels in the New Testament, and its Chinese gloss. In addition to the gloss, the Greek text is linked to two linguistic resources. First, it has been word-aligned to the Revised Chinese Union Version, the most recent Chinese translation of the New Testament; second, it has been annotated with word dependencies, adapted from dependency trees in the PROIEL project. Through a browser-based interface, one can perform bilingual string-based search on this interlinear text, possibly combined with dependency constraints. We have evaluated this interlinear with respect to the accuracy, consistency, and precision of its gloss, as well as its effectiveness as pedagogical material.

Hybrid Approach for Extracting Collocations from Arabic Quran Texts

Soyara Zaidi, Ahmed Abdelali, Fatiha Sadat and Mohamed-Tayeb Laskri

For the objective of building Quran Ontology using its original script, existing tools to develop such resource exists to support languages such as English or French hence can't be used for Arabic. In most cases, significant modifications must be made to obtain acceptable results. We present in this paper an automatic approach to extract simple terms and collocations to be used for the ontology. For extracting collocations, we use a hybrid approach that combines linguistic rule-based method, and Mutual-Information-based approach. We use a mutual information-based approach to filter, enhance and further improve the precision of the results obtained by linguistic method. Extracted collocations are considered essential domain terms to build Arabic ontology of Quran. We use The Crescent Quranic tagged Corpus which consisted of Quran text tagged per word, verse and chapter; it contains as well additional information about morphology and POS and syntax.

Semantic Relations-II.

Enhancing Resources and Applications

22 May 2012

ABSTRACTS

Editors:

Verginica Barbu Mititelu, Octavian Popescu, Viktor Pekar

Workshop Programme

09:00-10:30 – Semantic Relations Identification and Extraction

Sara Mendes, Silvia Necşulescu, Núria Bel, *Synonym Extraction Using a Language Graph Model*

Pilar León Araúz and Pamela Faber, *Causality in the Specialized Domain of the Environment*

Scott Piao, Diana Bental, Jon Whittle, Ruth Aylett, Stephann Makri, Xu Sun, *A Pilot Study: Deriving a Users' Goal Framework from a Corpus of Interviews and Diaries*

10:30 – 11:00 Coffee break

11:00 – 13:00 Enhancing Applications

Agata Cybulska and Piek Vossen, *Using Semantic Relations to Solve Event Coreference in Text*

Gerold Schneider, *Using Semantic Resources to Improve a Syntactic Dependency Parser*

Darja Fišer, Polona Gantar, Simon Krek, *Using Explicitly and Implicitly Encoded Semantic Relations to map Slovene WordNet and Slovene Lexical Database*

Sruti Rallapalli and Soma Paul, *Evaluating Scope for Labeling Nominal Compounds Using Ontology*

13:00 – 14:00 Lunch break

14:00 – 16:00 Enhancing Resources

Daniela Katunar, Matea Srebáčić, Ida Raffaelli, Krešimir Šojat, *Arguments for Phrasal Verbs in Croatian and Their Influence on Semantic Relations in Croatian WordNet*

Silke Scheible and Sabine Schulte im Walde, *Designing a Database of GermaNet-based Semantic Relation Pairs Involving Coherent Mini-Networks*

Vasile Rus, Mihai Lintean, Cristian Moldovan, William Baggett, Nobal Niraula, Brent Morgan, *The SIMILAR Corpus: A Resource to Foster the Qualitative Understanding of Semantic Similarity of Texts*

Ştefan Daniel Dumitrescu, *Building a Baseline Supervised Relation Extraction System Using Freely-Available Resources*

Abhimanu Kumar, Richard Chatwin, Joydeep Ghosh, *Simple Unsupervised Topic Discovery for Attribute Extraction in SEM Tasks using WordNet*

16:00 – 16:30 Coffee break

16:30 – 17:30 Invited Talk

Patrick Hanks, *Mapping Semantic Relations onto Patterns of Word Use using Corpus Evidence*

Workshop Organizers

Verginica Barbu Mititelu	RACAI
Octavian Popescu	FBK
Viktor Pekar	OUP

Workshop Programme Committee

Eduard Barbu	Universidad de Jaen
Antonio Branco	Faculdade de Ciências de Lisboa
Elena Cabrio	INRIA
Corina Forăscu	UAIC
Nuria Gala	LIF-CNRS
Patrick Hanks	UWE
Amaç Herdağdelen	Crimson Hexagon
Diana Inkpen	University of Ottawa
Radu Ion	RACAI
Elisabetta Jezek	Universita di Pavia
Svetla Koeva	IBL
Gerhard Kremer	Universität Heidelberg
Hristina Kukova	IBL
Claudia Kunze	Qualisys GmbH
Svetlozara Leseva	IBL
Bernanrdo Magnini	FBK
Emanuele Pianta	FBK
Reinhard Rapp	University of Leeds
Didier Schwab	Laboratoire d'Informatique de Grenoble
Carlo Strapparava	FBK
Sara Tonelli	FBK
Dan Tufiş	RACAI
Michael Zock	LIF-CNRS

Preface

The present edition of the workshop *Semantic Relations* builds on the interest manifested by the participants to the first workshop *Semantic Relations. Theory and Applications* (held in conjunction with LREC2010) and by the large scientific community of linguists and language engineers.

At the first edition, we aimed at bringing together researchers in computational linguistics and lexical semantics, discussing theoretical and practical aspects of semantic relations and answering the question of how computational linguists could benefit from the work done by theoretical linguists and vice versa. In this second edition we focus on the benefits recources development and practical tasks in Natural Language Processing (NLP) have from and for the studies in lexical semantics.

The experience accumulated in Corpus Linguistics has shown that, while a large part of language use is regular and predictable, there is a significant part of it that is highly irregular and ambiguous. The research in this area suggests that lexical knowledge at large, encompassing all the information about lexical units, as well as the relationships between them, is instrumental in the accurate processing of natural language by computational methods.

There is a rising interest, both in theoretical and computational linguistics, in investigating the types of information that must be represented, and how these types could be conveniently organized in the lexicon in order to adequately describe the process of semantic interpretation. Specifically, the focus is on the relevant information which renders explicit the combinatorial process through which the meaning is formed – at a first level, the process of composing the lexical meaning from morphological parts, and, at a second level, the process of combining the conceptual knowledge of individual words, such as ontological categories and various relationships among them, into phrases carrying definite syntactic and semantic structures.

At a morphological level two research lines are noticeable. On the one hand, the study of affixes by means of which new words are created from existing ones sheds light on certain semantic relations between stem and their derived words; these relations are valid at a cross-lingual level, thus transferable among aligned resources and usable in various application. On the one hand, there is a growing interest in research into semantic relations either within compounds or between simplex and compounds, in their semantic representation with benefits for NLP tasks. The interest goes even beyond compounds and extends to set phrases and terms, while various domains are favoured, especially biomedicine. While theoretical linguistics establishes the possible relations involving compouns, computational linguistics automatically predicts and classifies these relations.

At a phrase level, there is an ongoing effort in NLP to automatically extract a large spectrum of semantic relations from various (semi)structured texts. From IS-A, part and causality to person-affiliation, organization-location and to abstract patterns encoding the relationships between words in conventional usage, the semantic relations have made their way into natural language processing. The properties of semantic relations are exploited in the economical design of language resources: (i) the transitivity of hyponymy relation, for example, is appropriate for nouns hierarchical organization based on inheritance properties of natural language, (ii) the patterning of verbs behavior shows that it is possible to represent the interconnection between lexical knowledge and world knowledge in a computable way.

The information required by automatic text processing using lexical-semantic relations can be acquired through corpus investigation methods and through data analysis in a strong sense. Accordingly, the lexicon could and should be built in a bottom-up manner by validating the phenomena mined from corpora by various computational methods.

In this edition of the workshop we wanted to highlight the interrelation between the quality and coverage of resources and the quality of applications relying on semantic relations. Specifically, in the call for papers we solicited papers on the following topics:

- Knowledge representation and semantic relations
- Extraction of semantic relations from various sources
- Exploitation of semantic relations in NLP applications
- Co-occurrence and semantic relations
- Lexical knowledge, world knowledge and semantic representation
- Patterns and semantic relations
- Semantic relations and word formation (compounding and derivation)
- Semantic relations and language learning and acquisition
- Semantic relations and language generation
- Semantic relations and terminology
- WordNets development

Most of these topics lie at the heart of the papers that were accepted to the workshop.

We would like to thank all the authors who submitted papers, as well as the members of the Program Committee for the time and effort they contributed in reviewing the papers. We are grateful to prof. Patrick Hanks for accepting to give an invited talk.

The Editors

Semantic Relations Identification and Extraction

9:00-10:30
Chairperson: Verginica Barbu Mititelu

Synonym Extraction Using a Language Graph Model

Sara Mendes, Silvia Necşulescu, Núria Bel

One of the main requirements for lexical knowledge bases to be usable in NLP applications, apart from an appropriate data model, is a satisfactory level of coverage. Manually developed language resources are accurate, balanced and very reliable, but the cost of building them, both in terms of human resources and of time consumption has been a setback for their real application. Thus, conceiving methods for an automatic and fast development of language resources capable of providing adequate and reliable data for different languages and for different domains, if necessary, has become crucial in the field of NLP applications. The work presented here is framed by this general research effort. We aim at ultimately contributing to reduce the human effort required in the development of rich language resources in this work. We focus on the automatic acquisition of synonymy relations. We use a graph model and similarity measures based on the information encoded in the graph to extract lists of synonym pairs from corpus data, showing how semantic relations, specifically synonym relations, can be successfully extracted from corpus data in an automatic way.

Causality in the Specialized Domain of the Environment

Pilar León Araúz and Pamela Faber

EcoLexicon is a multilingual terminological knowledge base (TKB) that represents environmental concepts and their relations in different formats. In this paper we show how some of the manual processes that we have developed for the extraction and representation of semantic relations can be partially automatized with the help of NLP applications such as NooJ. Focusing on the causal relation, we have designed various graph-based micro-grammars to match and annotate the corpus. This permits the extraction of causal propositions, and identifies the terms that primarily act as causes and effects in environmental contexts. Finally, these grammars can also be used to measure the prototypicality of causal propositions within four different environmental domains.

A Pilot Study: Deriving a Users' Goal Framework from a Corpus of Interviews and Diaries

Scott Piao, Diana Bental, Jon Whittle, Ruth Aylett, Stephann Makri, Xu Sun

This paper describes pilot work in which we explore the feasibility of deriving a goal framework for the potential users of applications employing a grounded theory method based on a corpus of empirical data. The issue of developing and applying human goal frameworks has been studied in a number of areas, such as artificial intelligence and information seeking. But most existing goal frameworks are either constrained to a few information search related goals or mainly reflect highly abstract psychological motivations, and hence are not readily applicable to the applications which need to deal with complex practical users' goals. In this study, we employ corpus-based approach for goal framework development, and identify goal concepts and analyse semantic relations among them based on a collection of interview and diary transcripts. We suggest that our approach provides a feasible way of deriving goal frameworks for practical purposes as the corpus data tend to closely reflect the users' concrete requirements. Furthermore, our study reveals the need for more corpus resources for human goal analysis and automatic detection.

Enhancing Applications

11:00-13:00
Chairperson: Octavian Popescu

Using Semantic Relations to Solve Event Coreference in Text

Agata Cybulska and Piek Vossen

In this paper, we report on how semantic relations between event mentions in text can be used to solve event coreference. Event descriptions in text differ in specificity and granularity. We believe that based on meronymy and hyponymy relations between event mentions one can determine shifts in levels of granularity and abstraction and use these as indication for coreference resolution. This article presents a model that captures the relationship between semantic relations amongst events and event coreference. A number of heuristics can be used to estimate semantic distance between instances of event descriptions and based on that to calculate coreference match between event mentions. Within this study we used the Leacock-Chodorow similarity measure as a heuristic for event coreference resolution. We report about the success rates of our experiments based on the evaluation performed on a corpus annotated with coreferent events.

Using Semantic Resources to Improve a Syntactic Dependency Parser

Gerold Schneider

Probabilistic syntactic parsing has made rapid progress, but is reaching a performance ceiling. More semantic resources need to be included. We exploit a number of semantic resources to improve parsing accuracy of a dependency parser. We compare semantic lexica on this task, then we extend the back-off chain by punishing underspecified decisions. Further, a simple distributional semantics approach is tested. Selectional restrictions are employed to boost interpretations that are semantically plausible. We also show that self-training can improve parsing even without needing a re-ranker, as we can rely on a sufficiently good estimation of parsing accuracy. Parsing large amounts of data and using it in self-training allows us to learn world knowledge from the distribution of syntactic relation. We show that the performance of the parser considerably improves due to our extensions.

Using Explicitly and Implicitly Encoded Semantic Relations to map Slovene WordNet and Slovene Lexical Database

Darja Fišer, Polona Gantar, Simon Krek

In this paper we present the results of a case study in which we use explicitly and implicitly encoded semantic relations to automatically map lexical entries from two different lexical semantic resources for Slovene with the well-known Simplified Lesk algorithm. We explain the selection of the mapping sample and mapping elements and describe the pre-processing steps that were performed in order to facilitate the mapping procedure. Manual evaluation of the mappings shows promising results, especially for nouns which were correctly mapped in 68% of the cases. Discrepancies in the mappings are also analysed in order to gain insight into the conceptual differences between the resources and investigate possible future refinements of the mapping procedure.

Evaluating Scope for Labeling Nominal Compounds Using Ontology

Sruti Rallapalli and Soma Paul

Labeling nominal compounds with semantic relations is a challenging NLP task, as it requires the extraction of the hidden relation between the constituents of the nominal compound. In this paper, we explore the scope of identifying the semantic relation and thereby interpreting a nominal compound using an indexed, semantic ontology. This method has the following advantages over other approaches that use unstructured documents and classification models for nominal compound interpretation: 1. A semantic relation is much less ambiguous than a verb or preposition paraphrase. 2. Processing of an unstructured database is avoided. 3 Instances of infrequent nominal compounds are easier to handle, as there are no statistical predictions involved. However, one issue with our pro-posed system is the lack of robustness which arises due to the difficulty involved in obtaining a huge, generic ontology. This issue is addressed in our work by combining the ontology search with noun similarity measurement techniques to handle the cases that are not covered in our ontology.

Enhancing Resources

14:00-16:00
Chairperson: Verginica Barbu Mititelu

Arguments for Phrasal Verbs in Croatian and Their Influence on Semantic Relations in Croatian WordNet

Daniela Katunar, Matea Srebáčić, Ida Raffaelli, Krešimir Šojat

In this paper we introduce the category of phrasal verbs in Croatian lexicon and grammar description in order to show their influence on semantic relations, namely synonymy and polysemy in Croatian WordNet (henceforth CroWN). We discuss the practical and theoretical implications that arise from the introduction of the category of phrasal verbs in the description of the Croatian lexicon. We also address the interaction of synonymy and polysemy as manifested in the semantic relations of phrasal verbs to their monolexemic counterparts and facilitated by the structure of CroWN. The lemmatization of phrasal verbs in Croatian dictionaries and its modification for purposes of improving semantical relations in CroWN is also discussed. We also propose building of the Croatian phrasal verbs database, describe its structure and its further expanison which would facilitate extraction and incorporation of phrasal verbs into CroWN, and thus improve MT systems and information extraction via this computational lexical resource.

Designing a Database of GermaNet-based Semantic Relation Pairs Involving Coherent Mini-Networks

Silke Scheible and Sabine Schulte im Walde

We describe the design and compilation of a new database containing German semantic relation pairs drawn from the lexical network GermaNet. The database consists of two parts: A representative selection of lexical units drawn from the three major word classes adjectives, nouns, and verbs, which are balanced according to semantic category, polysemy, and type frequency (‘SemrelTargets’); and a set of semantically coherent GermaNet subnets consisting of semantic relations pairs clustering around the selected targets (‘SemrelNets’). The database, which contains 99 SemrelTargets for each of the three word classes, and a total of 1623 relation pairs distributed across the respective subnets, promises to be an important resource not only for research in computational linguistics, but also for studies in theoretical linguistics and psycholinguistics. Currently, the data is being used in two types of human judgement experiments, one focusing on the generation of semantically related word pairs, and the other on rating the strength of semantic relations.

The SIMILAR Corpus: A Resource to Foster the Qualitative Understanding of Semantic Similarity of Texts

Vasile Rus, Mihai Lintean, Cristian Moldovan, William Baggett, Nobal Niraula, Brent Morgan

We describe in this paper the SIMILAR corpus which was developed to foster a deeper and qualitative understanding of word-to-word semantic similarity metrics and their role on the more general problem of text-to-text semantic similarity. The SIMILAR corpus fills a gap in existing resources that are meant to support the development of text-to-text similarity methods based on word-level similarities. The existing resources, such as data sets annotated with paraphrase information between two sentences, do not provide word-to-word semantic similarity annotations and quality judgments at word-level. We annotated 700 pairs of sentences from the Microsoft Research Paraphrase corpus with word-to-word semantic similarity information using both a greedy and optimal protocol. We proposed a set of qualitative word-to-word semantic similarity relations which were then used to annotate the corpus. We also present a detailed analysis of various quantitative word-to-word semantic similarity metrics and how they relate to our qualitative relations. A software tool has been developed to facilitate the annotation of texts using the proposed protocol.

Building a Baseline Supervised Relation Extraction System Using Freely-Available Resources

Ştefan Daniel Dumitrescu

The article presents an easy-to-follow guide to building a supervised Relation Extraction system using free resources. The reader can see how to build the system in a step-by-step fashion, what tools, methods and data sources are needed and how they can be processed and then used, as well as see the practical results of such a system. Also, we explore the surface of performance evaluation giving the reader some basic measures and definitions, like: binary classifiers, cross-validation, feature space with different features extracted from annotated sentences, impact of features in different classifiers, confusion matrices and feature evaluation methods.

Simple Unsupervised Topic Discovery for Attribute Extraction in SEM Tasks using WordNet

Abhimanu Kumar, Richard Chatwin and Joydeep Ghosh

We present here a simple approach for topic discovery to extract attributes of online products using Wordnet. Identifying product attributes is important for search engine marketing (SEM) since it is integral to the ads displayed for search queries (Moran, 2009). Our wordnet based model provides a simple, scalable and high precision attribute extraction mechanism. It is well suited for identifying attributes for previously unseen product categories and thus works especially well for SEM scenario. It outperforms unsupervised topic discovery approaches such as LDA for SEM tasks on 4 online product datasets. The model has been successfully implemented as a production version code for ad-copy creation.

Invited talk

16:30-17:30

Chairperson: Octavian Popescu

Mapping Semantic Relations onto Patterns of Word Use using Corpus Evidence

Patrick Hanks

This paper presents an empirically well-founded corpus-driven theory of natural language as an analogical system of procedures governed by two interrelated sets of rules: rules for using language normally (idiomatically) and rules for exploiting those norms creatively. The theory is called the Theory of Norms and Exploitations (TNE). Examination of very large quantities of data for word use shows that words in isolation can present unresolvable problems of ambiguity, whereas phraseological patterns, which are not unlike the 'constructions' of construction grammar, are normally each associated with a unique meaning.

Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR

22 May 2012

ABSTRACTS

Editors:

Victoria Arranz, Daan Broeder, Bertrand Gaiffe, Maria Gavrilidou, Monica Monachini, Thorsten Trippel

Workshop Programme

22 May 2012

9:00 – 9:10 Welcome and Introduction

9:10 – 10:40 Overview Session

Daan Broeder, Menzo Windhouwer, Dieter van Uytvanck, Thorsten Trippel and Twan Goosen*,* *CMDI: a Component Metadata Infrastructure*

Maria Gavrilidou, Penny Labropoulou, Elina Desipri, Ioanna Giannopoulou, Olivier Hamon and Victoria Arranz, *The META-SHARE Metadata Schema: Principles, Features, Implementation and Conversion from other Schemas*

Gunn Inger Lyse, Carla Parra Escartín and Koenraad De Smedt, *Applying Current Metadata Initiatives: The META-NORD Experience*

10:40 – 11:00 Coffee break

11:00 – 13:00 Working and Experiencing the Component Model

Volker Boehlke, Torsten Compart and Thomas Eckart, *Building up a CLARIN resource center – Step 1: Providing metadata*

Thorsten Trippel, Christina Hoppermann and Griet Depoorter*, Infrastructure (CMDI) in a Project on Sustainable Linguistic Resources*

Hanna Hedeland and Kai Wörner, *Experiences and Problems creating a CMDI profile from an existing Metadata Schema*

Menzo Windhouwer, Daan Broeder and Dieter van Uytvanck, *A CMD Core Model for CLARIN Web Services*

13:00 – 14:00 Lunch break

14:00 – 16:00 General Impulses and Test Cases

Peter Menke and Philipp Cimiano, *Towards an ontology of categories for multimodal annotation*

Kristian Tangsgaard Hvelplund and Michael Carl, *User Activity Metadata for Reading, Writing and Translation Research*

Paula Estrella, *Metadata for a Mocoví - Quechua - Spanish parallel corpus*

Hennie Brugman and Mark Lindeman, *Publishing and Exploiting Vocabularies using the OpenSKOS Repository Service*

16:00 – 16:30 Coffee break

16:30 – 17:45 Metadata Applications: Overview Presentations and Demos

Peter Withers, *Metadata Management with Arbil*

Matej Durco, Daan Broeder and Menzo Windhouwer, *SMC4LRT - groundwork for query expansion and semantic search*

Martine de Bruin, Marc Kemps-Snijders, Jan Pieter Kunst, Maarten van der Peet, Rob Zeeman and Junte Zhang, *Applying CMDI in real life: the Meertens case*

Demos

17:45 – 18:30 Final Discussion and Wrap-up

Workshop Organizing Committee

Victoria Arranz	ELDA/ELRA, Paris, France
Daan Broeder	MPI, Nijmegen, The Netherlands
Bertrand Gaiffe	ATILF, Nancy, France
Maria Gavrilidou	ILSP/Athena R.C., Athens, Greece
Monica Monachini	CNR-ILC, Pisa, Italy
Thorsten Trippel	Universität Tübingen, Tübingen, Germany

Workshop Programme Committee

Helen Aristar-Dry	Michigan State University, USA
Núria Bel	UPF, Barcelona, Spain
António Branco	University of Lisbon, Portugal
Lars Borin	Språkbanken, Göteborg, Sweden
Khalid Choukri	ELDA/ELRA, Paris, France
Thierry Declerck	DFKI, Saarbrücken, Germany
Matej Durco	Austrian Academy of Sciences, Vienna, Austria
Gil Francopoulo	CNRS-LIMSI-IMMI + TAGMATICA, Paris, France
Francesca Frontini	CNR-ILC, Pisa, Italy
Olivier Hamon	ELDA/ELRA, Paris, France
Erhard Hinrichs	Universität Tübingen, Tübingen, Germany
Penny Labropoulou	ILSP/Athena R.C., Athens, Greece
Jan Odijk	Universiteit Utrecht, The Netherlands
Elena Pierazzo	Kings' College, London, UK
Laurent Romary	INRIA, Nancy, France
Andreas Witt	IDS, Mannheim, Germany
Peter Wittenburg	MPI, Nijmegen, The Netherlands
Tamás Varadi	Hungarian Academy of Sciences, Budapest, Hungary
Marta Villegas	UPF, Barcelona, Spain
Sue Ellen Wright	Kent State University, USA

Preface/Introduction

The description of Language Resources (LRs) continues to be a crucial point in the lifecycle of LRs, and more particularly, in their sustainable exchange. This has been so for a number of repositories or LR distribution centres in place (ELRA, GSK, LDC, OLAC, TST-Centrale, BAS, among others), who house LR catalogues following some proprietary metadata schema. A number of projects and initiatives have also focused these past few years in the sharing of LRs (ENABLER, CLARIN, FLaReNet, PANACEA, META-SHARE), for example, for Language Technology (LT).
Based on these initiatives a consensus emerges that shows a number of requirements for standardized metadata:

- There should be a common publication channel for the LR descriptions in the world.
- This channel allows users to carry out easy and efficient LR data discovery and possible subsequent retrieval of LRs.
- Expert knowledge is required to create the data model for the metadata description.
- Subject matter experts (both researchers and LR/LT providers and developers) are required to provide the content for the data model.
- The data model needs to be clear, expressive, flexible, customizable and interoperable.
- Metadata have to provide for different user groups, ranging from providers to consumers (both individuals and organisations). This applies both to the information contained in the metadata and the supporting tool infrastructure for creating, maintaining, distributing, harvesting and searching the metadata.

Currently several initiatives focus on metadata. From the realm of work done within initiatives like ENABLER and CLARIN descended the Component MetaData Infrastructure (CMDI, ISO TC 37 SC 4 work item for ISO 24622), which allows the combination of standard data categories (for example from ISO 12620, isocat.org) to components, which are combined into metadata profiles. Early versions of this model have been operational in repositories such as ELRA's, which complied with the work done within INTERA. FLaReNet, as the result of a permanent and cyclical consultation, has issued a set of main recommendations where a global infrastructure of uniform and interoperable metadata sets appears among the Top Priorities for the field of LRs. For use within HLT, META-SHARE provides a fully-fledged schema for the description of LRs, in the framework of the component model, covering all the current resource types and media types of use, in all the stages of a resource's lifecycle. Our aim is to learn from one another's experiences and plans in this area.

The current state of the art for metadata provision allows for a very flexible approach, catering for the needs of different archives and communities, referring to common data category registries that describe the meaning of a data category at least to authors of metadata. Component models for metadata provisions are for example used by CLARIN and META-SHARE, but there is also an increased flexibility in other metadata schemas such as Dublin Core, which is usually not seen as appropriate for meaningful description of language resources.

Making resources available for others and putting this to a second use in other projects has never been more widely accepted as a sensible efficient way to avoid a waste of efforts and resources. However, when it comes to the details, there is still a vast number of problems. This workshop has aimed at being a forum to address issues and challenges in the concrete work with metadata for LRs, not restricted to a single initiative for archiving LRs. It has allowed for exchange and discussion and we hope that the reader finds the articles here compiled interesting and useful.

Overview Session

Tuesday, 22 May, 9:10 – 10:40
Chairperson: Bertrand Gaiffe

CMDI: a Component Metadata Infrastructure

Daan Broeder, Menzo Windhouwer, Dieter van Uytvanck, Thorsten Trippel and Twan Goosen

The paper's purpose is to give an overview of the work on the Component Metadata Infrastructure (CMDI) that was implemented in the CLARIN research infrastructure. It explains the underlying schema, the accompanying tools and services. It also describes the status and impact of the CMDI developments done within the CLARIN project and past and future collaborations with other projects.

The META-SHARE Metadata Schema: Principles, Features, Implementation and Conversion from other Schemas

Maria Gavrilidou, Penny Labropoulou, Elina Desipri, Ioanna Giannopoulou, Olivier Hamon and Victoria Arranz

The current paper focuses on the presentation of a metadata model for the description of language resources proposed in the framework of the META-SHARE infrastructure, aiming to cover both datasets and tools/technologies used for their processing. It presents the rationale/background for its creation, the basic principles and features of the model, describes the process and results of its current application to the META-SHARE nodes, including the conversion from previous schemas, and concludes with work to be done in the future for the improvement of the model.

Applying Current Metadata Initiatives: The META-NORD Experience

Gunn Inger Lyse, Carla Parra Escartín and Koenraad De Smedt

In this paper we present the experiences with metadata in the Norwegian part of the META-NORD project, exemplifying the top-level description of language resources and tools (LRT). In recent years new initiatives have appeared as regards long-term accessibility plans to LRT. The META-NORD project, and the broader META-SHARE initiative in the META-NET network, are among the initiatives working on the standardization of the description of linguistic resources as well as on the creation of infrastructures that ensure a long-term curation and distribution of LRT. Here we present the use cases we have been dealing with in Norway as part of this effort. Furthermore, we also report on the importance of dealing with real user case scenarios to detect and solve potential problems concerning the construction of a larger open infrastructure for LRT.

Working and Experiencing the Component Model

Tuesday, 22 May 11:00 – 13:00
Chairperson: Monica Monachini

Building up a CLARIN resource center – Step 1: Providing metadata

Volker Boehlke, Torsten Compart and Thomas Eckart

In this paper we will describe the different problems that need to be solved in case one decides to provide metadata according to the CLARIN specifications for resource centers. We will report on why we decided to use the repository system fedora and how we configured it in order to serve our purposes. We will also describe how we designed CMDI components and profiles for our resources and how we dealt with the issues of granularity, updates/versioning of metadata. Additionally the usage of PIDs and PartIdentifiers will be discussed.

The Component Metadata Infrastructure (CMDI) in a Project on Sustainable Linguistic Resources
Thorsten Trippel, Christina Hoppermann and Griet Depoorter

The sustainable archiving of research data for predefined time spans has become increasingly important to researchers and is stipulated by funding organizations with the obligatory task of being observed by researchers. An important aspect in view of such a sustainable archiving of language resources is the creation of metadata, which can be used for describing, finding and citing resources. In the present paper, these aspects are dealt with from the perspectives of two projects: the German project for Sustainability of Linguistic Data at the University of Tübingen (NaLiDa) and the Dutch-Flemish HLT Agency hosted at the Institute for Dutch Lexicology (TST-Centrale). Both projects unfold their approaches to the creation of components and profiles using the Component Metadata Infrastructure (CMDI) as underlying metadata schema for resource descriptions, highlighting their experiences as well as advantages and disadvantages in using CMDI.

Experiences and Problems creating a CMDI profile from an existing Metadata Schema
Hanna Hedeland and Kai Wörner

To make language resources available through the CLARIN-D infrastructure, corpora of spoken discourse at the Hamburg Center for Language Corpora (Hamburger Zentrum für Sprachkorpora, HZSK) have to be described via CMDI compliant metadata. The aim is to create metadata that can be harvested automatically and can then be used in a federated search and browsing environment to facilitate discovery as well as recombination of existing resources. This paper describes the considerations, efforts and obstacles encountered in the process of creating a CMDI metadata profile for the HZSK. It had - based on an existing metadata format - to encompass most of the existing metadata, share as many existing components and profiles as possible and relate to metadata profiles that are being developed at other CLARIN-D projects that deal with similar resources.

A CMD Core Model for CLARIN Web Services
Menzo Windhouwer, Daan Broeder and Dieter van Uytvanck

In the CLARIN infrastructure various national projects have started initiatives to allow users of the infrastructure to create chains or workflows of web services. The Component Metadata (CMD) core model for web services described in this paper tries to align the metadata descriptions of these various initiatives. This should allow chaining/workflow engines to find matching and invoke services. The paper described the landscape of web services architectures and the state of the national initiatives. Based on this a CMD core model for CLARIN is proposed, which by using, within some limits, the standard facilities of the CMD Infrastructure can be adapted to the specific needs of a specific web service initiative. The paper closes with the current state and usage of the model and a look into the future.

General Impulses and Testcases

Tuesday, 22 May, 14:00 – 16:00
Chairperson: Daan Broeder

Towards an ontology of categories for multimodal annotation

Peter Menke and Philipp Cimiano

We examine how multimodal data collections, resulting mainly from (psycho)linguistic experiments, can be expressed in standardized metadata description formats. We summarize how such data collections differ structurally from the traditional concept of corpora, and we list thoughts, problems, and solutions that occurred when we designed and collected ISOcat data categories and CMDI components for the metadata representation of these data collections. As a result we present plans for an ontology of modalities and related concepts and data units, which we consider a more appropriate environment for the kind of multimodal data we are dealing with.

User Activity Metadata for Reading, Writing and Translation Research

Kristian Tangsgaard Hvelplund and Michael Carl

While there exists a large amount of static linguistic resources together with annotation schema and metadata, not much work has been done to describe the processes by which texts are produced. In the mid-1980s, translation process research began to use advanced technologies such as keyboard logging and more recently eye-tracking and screen recording to record and study user activity data of human reading, writing and translation processes. There has not however been much effort to synchronize and annotate the collected data. This paper suggests a structure for these processes along four dimensions. As the process data depends not only on the human writer/translator, but also on the type of text to be produced and the purpose of the final product, this paper suggests a metadata structure for user activity data which accounts for these different dimensions.

Metadata for a Mocoví - Quechua - Spanish parallel corpus

Paula Estrella

In this paper we present the work done to create the metadata associated to a parallel corpus of endangered languages spoken in Argentina, namely Mocoví and Quechua. Creating metadata is of great importance not only to document the resource and the language but also to make it available to the general public though browse and search facilities, given that resources for Amerindian languages are so few and so difficult to find. However, choosing an appropriate schema is not a trivial task if compatibility and interoperability are in mind. Therefore, it was decided to reuse previous work by major initiatives in language archiving and documentation, resulting in the customized IMDI schema described in this article.

Publishing and Exploiting Vocabularies using the OpenSKOS Repository Service

Hennie Brugman and Mark Lindeman

Many vocabularies in eHumanities and eCulture domains can, and increasingly often are converted to SKOS. The OpenSKOS web service platform provides easy ways to publish, upload, update, harvest, query and distribute SKOS vocabulary data. This has benefits for vocabulary builders, vocabulary consumers and builders of tools that exploit vocabularies. In this paper we present and discuss the OpenSKOS system and a number of its applications, including an application from the domain of linguistic resources and tools.

Metadata Applications: Overview Presentations and Demos

Tuesday, 22 May, 16:30 – 17:45 (Overviews 16:30-17:15, Demos in parallel afterwards)
Chairperson: Maria Gavrilidou

Metadata Management with Arbil

Peter Withers

Arbil is an application designed to create and manage metadata for research data and to arrange this data into a structure appropriate for archiving. The metadata is displayed in tables, which allows an overview of the metadata and the ability to populate and update many metadata sections in bulk. Both IMDI and Clarin metadata formats are supported and Arbil has been designed as a local application so that it can also be used offline, for instance in remote field sites. The metadata can be entered in any order or at any stage that the user is able; once the metadata and its data are ready for archiving and an Internet connection is available it can be exported from Arbil and in the case of IMDI it can then be transferred to the main archive via LAMUS (archive management and upload system).

SMC4LRT - groundwork for query expansion and semantic search

Matej Durco, Daan Broeder and Menzo Windhouwer

This paper describes a module of the Component Metadata Infrastructure, that allows query expansion by providing mappings between search indexes. This enables semantic search, ultimately increasing the recall when searching in metadata collections. The module builds on the Data Category Registry and Component Metadata Framework that are part of CMDI.

Applying CMDI in real life: the Meertens case

Martine de Bruin, Marc Kemps-Snijders, Jan Pieter Kunst, Maarten van der Peet, Rob Zeeman and Junte Zhang

The CMDI (Component Metadata Infrastructure) has gained widespread acceptance across multiple projects and organizations. To incorporate this approach many organizations need to adjust their organizational and technological structure to unlock the potential of the CMDI approach. The Meertens Institute has applied the CMDI approach to a large number of projects covering the full life cycle of the CMDI process, including metadata creation, ingest, publication and search processes. This paper covers our experiences with the CMDI approach and describes various aspects of our work process and projects in which the CMDI approach was adopted.

Final Discussion

Tuesday, 22 May, 17:45 – 18:30
Chairperson: Victoria Arranz and Thorsten Trippel

META-RESEARCH Workshop on Advanced Treebanking

22 May 2012

ABSTRACTS

Editors:

Jan Hajič, Koenraad De Smedt, Marko Tadić, António Branco

Workshop Programme

22 May 2012

09:00 – 10:30 Oral presentations – Session I
co-chaired by Jan Hajič and Koenraad De Smedt

09:00 – 09:10 Jan Hajič *Welcome and Introduction to the Workshop*

09:10 – 09:35 Tom Vanallemeersch *Parser-independent Semantic Tree Alignment*

09:35 – 10:00 Philippe Blache and Stéphane Rauzy
Hybridization and Treebank Enrichment with Constraint-Based Representations

10:00 – 10:25 Bruno Guillaume and Guy Perrier
Semantic Annotation of the French Treebank with Modular Graph Rewriting

10:25 – 10:30 Jan Hajič *Introduction to the Poster Session*

10:30 – 11:30 Coffee break with poster presentations

Victoria Rosén, Koenraad De Smedt, Paul Meurer and Helge Dyvik
An Open Infrastructure for Advanced Treebanking (with demo)

Oleg Kapanadze *A German-Georgian Parallel Treebank Project*

Tomáš Jelínek, Vladimír Petkevič, Alexandr Rosen and Hana Skoumalová
Czech Treebanking Unlimited

João Silva, António Branco, Sérgio Castro and Francisco Costa
Deep, Consistent and also Useful: Extracting Vistas from Deep Corpora for Shallower Tasks

11:30 – 13:00 Oral presentations and invited speech – Session II
co-chaired by Marko Tadić and António Branco

11:30 – 11:45 Hans Uszkoreit, coordinator of META-NET (invited speech)
High-Quality Research, New Language Resources and their Sharing

11:45 – 12:10 Rajesh Bhatt and Fei Xia
Challenges in Converting between Treebanks: a Case Study from the HUTB

12:10 – 12:35 Maytham Alabbas and Allan Ramsay
Arabic Treebank: from Phrase-Structure Trees to Dependency Trees

12:35 – 13:00 Gyri Losnegaard, Gunn Inger Lyse, Martha Thunes, Victoria Rosén, Koenraad De Smedt, Helge Dyvik and Paul Meurer
What We Have Learned from Sofie: Extending Lexical and Grammatical Coverage in an LFG Parsebank

13:00 End of Workshop (start of lunch break)

Workshop Organizers

Jan Hajič	Charles University in Prague, Czech Republic
Koenraad De Smedt	University of Bergen, Norway
Marko Tadić	University of Zagreb, Croatia
António Branco	University of Lisbon, Portugal

Workshop Programme Committee

António Branco	University of Lisbon, Portugal
Sabine Buchholz	Toshiba Research Europe, Cambridge, UK
Khalid Choukri	ELRA/ELDA, Paris, France
Silvie Cinková	Charles University in Prague, Czech Republic
Dan Cristea	University of Iaşi, Romania
Koenraad De Smedt	University of Bergen, Norway
Rebecca Dridan	University of Oslo, Norway
Nancy Ide	Vassar College, New York, USA
Valia Kordoni	DFKI, Berlin, Germany
Sandra Kuebler	Indiana University, Bloomington, USA
Krister Lindén	University of Helsinki, Finland
Paul Meurer	Uni Computing/Uni Research, Bergen, Norway
Adam Meyers	New York University, USA
Joakim Nivre	University of Uppsala, Sweden
Stephan Oepen	University of Oslo, Norway
Marco Passarotti	Catholic University of the Sacred Heart, Milan, It.
Eiríkur Rögnvaldsson	University of Reykjavik, Iceland
Victoria Rosén	University of Bergen, Norway
Mária Šimková	Slovak Academy of Sciences, Bratislava, Slovakia
Barbora Vidová Hladká	Charles University in Prague, Czech Republic
Fei Xia	University of Washington, USA
Daniel Zeman	Charles University in Prague, Czech Republic

Preface

Many R&D projects and research groups are creating, standardizing, converting and/or using treebanks, thereby often tackling the same issues and reinventing methods and tools. While a fair amount of treebanks have been produced in recent years, it is still a challenge for researchers and developers to reuse treebanks in suitable formats for new purposes. Standardization of interchange formats, conversion and adaptation to different purposes, exploration with suitable tools, long term archiving and cataloguing, and other issues still require significant efforts.

In this spirit, the present workshop has been conceived by four projects, namely T4ME, META-NORD, CESAR and META4U, which under the META-NET umbrella project strive to make many treebanks and other language resources and tools available for R&D. It is hoped that the workshop will contribute to innovative insights that will promote development, dissemination, use and reuse of treebanks in the future.

Thirteen papers were submitted to the workshop, of which ten were accepted for presentation at this half-day workshop. Six were selected for oral presentation while four were selected for poster presentation. We thank all our reviewers for their constructive evaluation of the papers.

Jan Hajič
Koenraad De Smedt
Marko Tadić
António Branco

Session I

Tuesday 22 May, 9:00 – 9:10
Chairperson: Koenraad De Smedt

Welcome and Introduction to the META-RESEARCH: Workhop on Advanced Treebanking

Jan Hajič

Session I

Tuesday 22 May, 9:10 – 9:35
Chairperson: Koenraad De Smedt

Parser-independent Semantic Tree Alignment

Tom Vanallemeersch

We describe an approach for training a semantic role labeler through cross-lingual projection between different types of parse trees, with the purpose of enhancing tree alignment on the level of syntactic translation divergences. After applying an existing semantic role labeler to parse trees in a resource-rich language (English), we partially project the semantic information to the parse trees of the corresponding target sentences, based on word alignment. After this precision-oriented projection, we apply a method for training a semantic role labeler which consists in determining a large set of features describing target predicates, roles and predicate-role connections, independently from the type of tree annotation (phrase structure or dependencies). These features describe tree paths starting at or connecting nodes. The semantic role labeling method does not require any knowledge of the parser nor manual intervention. We evaluated the performance of the cross-lingual projection and semantic role labeling using an English parser assigning PropBank labels and Dutch manually annotated parses, and are currently studying ways to use the predicted semantic information for enhancing tree alignment.

Session I

Tuesday 22 May, 9:35 – 10:00
Chairperson: Jan Hajič

Hybridization and Treebank Enrichment with Constraint-Based Representations

Philippe Blache and Stéphane Rauzy

We present in this paper a method for hybridizing constituency treebanks with constraint-based descriptions and enrich them with an evaluation of sentence grammaticality. Such information is calculated thanks to a two-steps technique consisting in: (1) constraint grammar induction from the source treebank and (2) constraint evaluation for all sentences, on top of which a grammaticality index is calculated. This method is theoretically-neutral and language independent. Because of the precision of the encoded information, such enrichment is helpful in different perspectives, for example when designing psycholinguistics experiments such as comprehension or reading difficulty.

Session I

Tuesday 22 May, 10:00 – 10:25
Chairperson: Jan Hajič

Semantic Annotation of the French Treebank with Modular Graph Rewriting

Bruno Guillaume and Guy Perrier

We propose to annotate the French Treebank with semantic dependencies in the framework of DMRS starting from an annotation with surface syntactic dependencies and using modular graph rewriting. This system has been experimented on the whole French Treebank with the prototype which implements the rewriting calculus.

Poster Session Introduction

Tuesday 22 May, 10:25 – 10:30
Chairperson: Jan Hajič

Introduction to the Poster Session

Poster Session (Coffee Break)

Tuesday 22 May, 10:30 – 11.30

An Open Infrastructure for Advanced Treebanking (with demo)

Victoria Rosén, Koenraad De Smedt, Paul Meurer and Helge Dyvik

Increases in the number and size of treebanks, and the complexity of their annotation, present challenges to their exploration by the research community. Adhering to different formalisms, lacking clear standards, and requiring specialized search and visualization and other services, treebanks have not been widely accessible to a broad audience and have remained underexploited. The INESS project is providing the first infrastructure integrating treebank annotation, analysis and distribution, bringing together treebanks for many different languages, spanning different annotation schemes and including parallel treebanks. The infrastructure offers a uniform interface, interactive visualizations, leading edge search capabilities and high performance computing.

Poster Session (Coffee Break)

Tuesday 22 May, 10:30 – 11.30

A German-Georgian Parallel Treebank Project

Oleg Kapanadze

This poster reports about efforts on building a parallel treebank for a typologically dissimilar language pair, namely German and Georgian. The project aims at supporting interdisciplinary collaboration in the field of jurisprudence adding a Natural Language Technology (NLT) angle to the human translation issue. The objective of this project is development of a bilingual Treebank which will be based on the German-Georgian parallel legislative texts.

Poster Session (Coffee Break)

Tuesday 22 May, 10:30 – 11.30

Czech Treebanking Unlimited

Tomáš Jelínek, Vladimír Petkevič, Alexandr Rosen and Hana Skoumalová

We build a large treebank of Czech, avoiding manual effort by using existing parsers, supplemented by a rule-based correction tool. A potentially underspecified morphological and syntactic annotation scheme offers multiple visualisation and export options, customisable in shape and detail according to the preferences of humans or computer applications. The annotation scheme consists of three layers: graphemics, morphology and constituency-based syntax, and is supported by lexicon (with a morphological, multi-word and syntactic part) and grammar. Annotation on any of the interlinked layers can be missing; ambiguous or undecidable phenomena are represented by underspecification and distributive disjunction.

Poster Session (Coffee Break)

Tuesday 22 May, 10:30 – 11.30

Deep, Consistent and also Useful: Extracting Vistas from Deep Corpora for Shallower Tasks

João Silva, António Branco, Sérgio Castro and Francisco Costa

Annotated corpora are fundamental for NLP, and the trend in their development is to move towards datasets with increasingly detailed linguistic annotation. To cope with the complexity of producing such resources, some approaches rely on a supporting deep processing grammar that provides annotation that is rich and consistent over its morphological, syntactic and semantic layers. However, for some purposes, the deep linguistic corpora thus produced are "too deep" and unwieldy. For instance, if one wishes to obtain a probabilistic constituency parser by learning a model over a treebank, the full extent of the annotation created by a deep grammar is not needed and can even be detrimental to training. In this poster, we report on procedures that, starting from a deep dataset produced by a deep processing grammar, extract a variety of vistas---that is, subsets of the information contained in the full dataset. This allows to take a single base dataset as a starting point and deliver a variety of corpora that are more streamlined and focused on particular tasks.

Session II

Tuesday 22 May, 11:30 – 11:45
Chairperson: Marko Tadić

High-Quality Research, New Language Resources and their Sharing

Invited speech

Hans Uszkoreit, main coordinator of META-NET

Session II

Tuesday 22 May, 11:45 – 12:10
Chairperson: Marko Tadić

Challenges in Converting between Treebanks: a Case Study from the HUTB

Rajesh Bhatt and Fei Xia

An important question for treebank development is whether high-quality conversion from one representation (e.g., dependency structure) to another representation (e.g., phrase structure) is possible, assuming that annotation guidelines exist for both representations. In this study, we demonstrate that the conversion is possible only under certain conditions, and even when the conditions are met, the conversion is complex as we need to examine the two sets of guidelines on a phenomenon-by-phenomenon basis and provide an intermediate representation for phenomena with incompatible analysis.

Session II

Tuesday 22 May, 12:10 – 12:35
Chairperson: António Branco

Arabic Treebank: from Phrase-Structure Trees to Dependency Trees

Maytham Alabbas and Allan Ramsay

The aim here is to create a dependency treebank from a phrase-structure treebank for Arabic. Arabic has a number of characteristics, described below, which make it particularly challenging to any natural language processing (NLP) applications. We describe an encouraging semi-automatic technique for converting phrase-structure trees to dependency trees by using a head percolation table. One of the most significant challenges here is the determination of the head of each subtree. We therefore examined different versions of the head percolation table to find the best priority list for each entry in the table. Given that there is no absolute measure of the 'correctness' of a conversion of a phrase structure tree to dependency form, we tested the various transformations by seeing how well a state-of-the-art dependency parser learnt the generalisations that were embodied by the converted trees.

Session II

Tuesday 22 May, 12:35 – 13:00
Chairperson: António Branco

What We Have Learned from Sofie: Extending Lexical and Grammatical Coverage in an LFG Parsebank

Gyri Losnegaard, Gunn Inger Lyse, Martha Thunes, Victoria Rosén, Koenraad De Smedt, Helge Dyvik and Paul Meurer

Constructing a treebank as a dynamically parsed corpus is an iterative process which may effectively lead to improvements of the grammar and lexicon. We show this from our experiences with semiautomatic disambiguation of a Norwegian LFG parsebank. The main types of grammar and lexicon changes necessary for achieving improved coverage are analyzed and discussed. We show that an important contributing factor to missing coverage is missing multiword expressions in the lexicon.

Adaptation of Language Resources and Tools for Processing Cultural Heritage Objects

26 May 2012

ABSTRACTS

Editors:

Petya Osenova, Stelios Piperidis, Milena Slavcheva, Cristina Vertan

Workshop Programme

09:00 – 09:30 – Introduction by Workshop Organisers

09:30 – 10:30 Invited talk: Stefanie Diepper. *Automatic methods for historical language data: studies on rule-based normalization, part-of-speech and morphological tagging*

10:30 – 11:00 Coffee break

11:00 – 11:20 Mikhail Gronas, Anna Rumshisky, Aleksandar Gabrovski, Samuel Kovaka and Hongyu Chen. *Tracking the history of knowledge using historical editions of Encyclopedia Britannica*

11:20 – 11:40 Jirka Hana, Boris Lehečka, Anna Feldman, Alena Černá and Karel Oliva. *Building a corpus of Old Czech*

11:40 – 12:00 Agnieszka Mykowiecka, Piotr Rychlik and Jakub Waszczuk. *Building an electronic dictionary of Old Polish on the base of the paper resource*

12:00 – 12:20 Serge Heiden and Alexei Lavrentiev. *The TXM portal software giving access to Old French manuscripts online*

12:20 – 12:40 Lauma Pretkalnina, Peteris Paikens, Normunds Gruzitis, Laura Rituma and Andrejs Spektors. *Making historical Latvian texts more intelligible to contemporary readers*

12:40 – 13:00 Baldev Ram Khandoliyan, Rajneesh Kumar Pandey, Archana Tiwari and Girish Nath Jha. *Text encoding and search for Āyurvedic texts: an interconnected lexical database*

13:00 – 13:20 Tobias Sippel and Jan-Torsten Milde. *A multitouch enabled annotation editor for digitized historical documents*

13:20 – 13:40 Dolf Trieschnigg, Djoerd Hiemstra, Mariët Theune, Franciska de Jong and Theo Meder. *An exploration of language identification techniques for the Dutch Folktale Database*

13:40 – 13:45 Closing session

Workshop Organisers

Petya Osenova	St. Kl. Ohridski University of Sofia and IICT, Bulgarian Academy of Sciences, Bulgaria
Stelios Piperidis	Institute for Language and Speech Processing, Athens, Greece
Milena Slavcheva	IICT, Bulgarian Academy of Sciences, Bulgaria
Cristina Vertan	University of Hamburg, Germany

Workshop Programme Committee

David Baumann	School of Computer Science, Carnegie Mellon, USA
Walter Daelemans	University of Antwerp, Belgium
Günther Görz	University Erlangen, Germany
Walther v. Hahn	University of Hamburg, Germany
Piroska Lendvai	Hungarian Academy of Sciences, Hungary
Anke Lüdeling	Humboldt University, Berlin, Germany
Gábor Prószéky	MorphoLogic, Hungary
Laurent Romary	LORIA-INRIA, Nancy, France
Éric Laporte	Université Paris-Est Marne-la-Vallée, France
Kiril Simov	IICT, Bulgarian Academy of Sciences, Bulgaria
Manfred Thaler	Cologne University, Germany
Tamás Váradi	Hungarian Academy of Sciences, Hungary
Martin Wynne	University of Oxford, U.K.
Kalliopi Zervanou	University of Tilburg, The Netherlands

Preface

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 in the form of Digital Libraries.
Most parts of these libraries are made available not only to researchers in a certain Humanities domain (such as, 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 Language Technology 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.
- In most cases, the historical variants of the language lack firmly established syntactic or morphological structures and that makes the definition of a robust set of rules extremely difficult. Historical texts often constitute a mixture of several languages including Latin, Old Greek, Slavonic, etc.
- Historical texts contain a great number of abbreviations, which follow different models.
- The conception of the world is somewhat different from ours (that is, different thinking about the Earth, different views in medicine, astronomy, etc.), which makes it more difficult to build the necessary knowledge bases.

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.

The current workshop tries to address those issues. The proceedings contains eight papers dealing with historical variants of languages such as French, Dutch, German, Czech, Polish, Latvian, Sanscrit. The paper topics range from the creation of language resources and their intelligent representation for non-specialist users to the development of automatic tools for processing historical language variants.

We would like to thank all contributors, our invited speaker, and especially the members of the programme committee who completed the review process in extremely short time.

The Organisers

Invited Talk

Saturday 26 May, 9:30 – 10:30

Chairperson:

Automatic Methods for Historical Language Data: Studies on Rule-Based Normalization, Part-of-Speech and Morphological Tagging

Stefanie Diepper

Analysis of historical languages differs from that of modern languages in two important points. First, there are no agreed-upon, standardized writing conventions. Instead, characters and symbols used by the writer of some manuscript in parts reflect impacts as different as spatial constraints (parchment is expensive and, hence, use of abbreviations seems favorable) or dialect influences (the dialect spoken by the author of the text, or the writer's dialect, who writes up or copies the text, or even the dialect spoken by the expected readership). This often leads to inconsistent spellings, even within one text written up by one writer. Second, resources of historical languages are scarce and often not very voluminous.

These features - variance in the data and lack of large resources - challenge many statistical analysis tools, whose quality usually depend on the availability of large training samples. A common way to tackle these problems is by normalizing historical spellings, by mapping them to modern wordforms or some virtual historical standardized forms.

In the talk, I will present an unsupervised, rule-based approach to wordform normalization. Rules are specified in the form of context-aware rewrite rules that apply to sequences of characters. The rules are derived from two aligned versions of the Luther bible. I will also present results from a set of tagging experiments. In these experiments, a state-of-the-art tagger is applied to original and normalized wordforms, to assign part-of-speech and morphological tags. The data used in this research are texts from Middle and Early New High German.

Session Regular Papers

Saturday 26 May, 11:00 – 13:45

Chairperson: Milena Slavcheva

Tracking the History of Knowledge Using Historical Editions of Encyclopedia Britannica

Mikhail Gronas, Anna Rumshisky, Aleksandar Gabrovski, Samuel Kovaka and Hongyu Chen

Despite the wealth of newly available digital materials, the scope of text-based investigations has mostly been limited to either synchronous or short-term historical analysis. In this paper, we report on the first stage of the project that focuses on tracking long-range historical change, specifically, on the history of ideas and concepts. The project's aim is to map out the history of representation of knowledge in Europe over last three centuries using as a proxy the history of changes in historical editions of Encyclopedia Britannica. We describe a series of corpus-analytical tasks necessary for building the analytical and comparative tools for historical analysis using scanned noisy text. In this first stage of the project, we focus specifically on the tools for tracking and visualizing the relative importance of people, interconnections between them, and the rise and fall of their reputations.

Building a Corpus of Old Czech

Jirka Hana, Boris Lehečka, Anna Feldman, Alena Černá and Karel Oliva

In this paper we describe our efforts to build a corpus of Old Czech. We report on tools, resources and methodologies used during the corpus development as well as discuss the corpus sources and structure, the tagset used, the approach to lemmatization, morphological analysis and tagging. Due to practical restrictions we adapt resources and tools developed for Modern Czech. However, some of the described challenges, such as the non-standardized spelling in early Czech and the form and lemma variability due to language change during the covered time-span, are unique and never arise when building synchronic corpora of Modern Czech.

Building an Electronic Dictionary of Old Polish on the Base of the Paper Resource

Agnieszka Mykowiecka, Piotr Rychlik and Jakub Waszczuk

In this paper we present a process of converting an existing dictionary of Old Polish into LMF (Lexical Markup Framework) format. We discuss problems related to the transformation of a resource build to be used as a paper book into its electronic version. We describe the subsequent stages of the process consisting in scanning the paper source followed by OCR and correction of its results, converting the data into LMF format and enriching the dictionary with information which was given indirectly or which can be obtained from other sources. In particular we describe the method of assigning part of speech names to lexicon entries.

The TXM Portal Software Giving Access to Old French Manuscripts Online

Serge Heiden and Alexei Lavrentiev

This paper presents the new TXM software platform giving online access to Old French Text Manuscripts images and tagged transcriptions for concordancing and text mining. This platform is able to import medieval sources encoded in XML according to the TEI Guidelines for linking manuscript images to transcriptions, encode several diplomatic levels of transcription including abbreviations and word level corrections. It includes a sophisticated tokenizer able to deal with TEI tags at different levels of linguistic hierarchy. Words are tagged on the fly during the import process using IMS TreeTagger tool with a specific language model. Synoptic editions displaying side by side manuscript images and text transcriptions are automatically produced during the import process. Texts are organized in a corpus with their own metadata (title, author, date, genre, etc.) and several word properties indexes are produced for the CQP search engine to allow efficient word patterns search to build different type of frequency lists or concordances. For syntactically annotated texts, special indexes are produced for the Tiger Search engine to allow efficient syntactic concordances building. The platform has also been tested on classical Latin, ancient Greek, Old Slavonic and Old Hieroglyphic Egyptian corpora (including various types of encoding and annotations).

Making Historical Latvian Texts More Intelligible to Contemporary Readers

Lauma Pretkalnina, Peteris Paikens, Normunds Gruzitis, Laura Rituma and Andrejs Spektors

In this paper we describe an ongoing work developing a system (a set of web-services) for transliterating the Gothic-based Fraktur script of historical Latvian to the Latin-based script of contemporary Latvian. Currently the system consists of two main components: a generic transliteration engine that can be customized with alternative sets of rules, and a wide coverage explanatory dictionary of Latvian. The transliteration service also deals with correction of typical OCR errors and uses a morphological analyzer of contemporary Latvian to acquire lemmas – potential headwords in the dictionary. The system is being developed for the National Library of Latvia in order to support advanced reading aids in the web-interfaces of their digital collections.

Text encoding and search for Āyurvedic texts: An interconnected lexical database

Baldev Ram Khandoliyan, Rajneesh Kumar Pandey, Archana Tiwari and Girish Nath Jha

The paper explores an interconnected lexical resource system for key texts of Āyurveda. The system which is in the form of an online index can be tested at http://sanskrit.jnu.ac.in/ayur/index.jsp. The paper also discusses the text encoding mechanisms and search processes that have been used to create the resource. Though Āyurvedic texts which have similar structure. Āyurveda has had a long tradition of texts and commentaries, we have taken only two key texts - Suśruta Samhitā and Caraka Samhitā and a glossary called Bhāvaprakāśa Nighaṇṭu with Amarakośā. The system works as an interactive and multi-dimensional knowledge based indexing system with search facility for these mainstream Āyurvedic texts and has potentials for use as a generic system for all.

A Multitouch Enabled Annotation Editor for Digitized Historical Documents

Tobias Sippel and Jan-Torsten Milde

In this paper we describe a system allowing to annotate digitized historical documents stored in METS/MODS format. The annoation is spacially connected to the original document. Both grafical and textual annoation are possible. The system is equipped with an intuitive touch based control and is designed to be a simple to use workbench for scientists working with historical texts.

An Exploration of Language Identification Techniques for the Dutch Folktale Database

Dolf Trieschnigg, Djoerd Hiemstra, Mariët Theune, Franciska de Jong and Theo Meder

The Dutch Folktale Database contains fairy tales, traditional legends, urban legends, and jokes written in a large variety and combination of languages including (Middle and 17th century) Dutch, Frisian and a number of Dutch dialects. In this work we compare a number of approaches to automatic language identification for this collection. We show that in comparison to typical language identification tasks, classification performance for highly similar languages with little training data is low. The studied dataset consisting of over 39,000 documents in 16 languages and dialects is available on request for followup research.

Third Workshop on Building and Evaluating Resources for Biomedical Text Mining

26 May 2012

ABSTRACTS

Editors:

Sophia Ananiadou, Kevin Cohen, Dina Demner-Fushman, Paul Thompson

Workshop Programme

09:15 – 09:30 – Welcome (Sophia Ananiadou)

09:30 – 10:30 - Invited Talk (chair: Kevin Cohen)

Jun'ichi Tsujii, Microsoft Research Asia
Semantic and linguistic annotations in GENIA

10:30 – 11:00 - Coffee break

11:00 – 12:15 - Session 1 (chair: Paul Thompson)

11:00 – 11:25
Claudiu Mihăilă, Riza Theresa Batista-Navarro and Sophia Ananiadou
Analysing Entity Type Variation across Biomedical Subdomains

11:25 – 11:50
Suwisa Kaewphan, Sanna Kreula, Sofie Van Landeghem, Yves Van de Peer, Patrik R. Jones and Filip Ginter
Integrating Large-Scale Text Mining and Co-Expression Networks: Targeting NADP(H) Metabolism in E. coli with Event Extraction

11:50 – 12:15
Mariana Neves, Alexander Damaschun, Andreas Kurtz and Ulf Leser
Annotating and Evaluating Text for Stem Cell Research

12:15 – 14:00 Lunch break

14:00 – 15:15 - Session 2 (chair: Kevin Cohen)

14:00 – 14:25
Raheel Nawaz, Paul Thompson and Sophia Ananiadou
Meta-Knowledge Annotation at the Event Level: Comparison between Abstracts and Full Papers

14:25 – 14:50
Fei Xia and Meliha Yetisgen-Yildiz
Clinical Corpus Annotation: Challenges and Strategies

14:50- 15:15
Dimitrios Kokkinakis
The Journal of the Swedish Medical Association - a Corpus Resource for Biomedical Text Mining in Swedish

15:15– 15:45 - Session 3 – Short poster presentations (chair: Paul Thompson)

15:15 – 15:25
Hercules Dalianis and Henrik Boström
Releasing a Swedish Clinical Corpus after Removing All Words - De-Identification Experiments with Conditional Random Fields and Random Forests

15:25- 15:35
Alyaa Alfalahi, Sara Brissman and Hercules Dalianis
Pseudonymisation of Personal Names and other PHIs in an Annotated Clinical Swedish Corpus

15:35 – 15:45
Olfa Makkaoui, Julien Desclés and Jean-Pierre Desclés
Evaluation and Performance Improvement of the BioExcom System for the Automatic Detection of Speculation in Biomedical Texts

15:45 – 16:30 – Poster session and coffee break

16:30 – 17:20 - Session 4 (chair: Sophia Ananiadou)

16:30 – 16:55
Philippe Thomas, Tamara Bobić, Martin Hofmann-Apitius, Ulf Leser and Roman Klinger
Weakly Labeled Corpora as Silver Standard for Drug-Drug and Protein-Protein Interaction

16:55 – 17:20
Amber Stubbs
Developing Specifications for Light Annotation Tasks in the Biomedical Domain

17:20 – 17:30 - Concluding remarks (chair: Sophia Ananiadou)

Workshop Organizers

Sophia Ananiadou	University of Manchester, UK
Kevin Cohen	University of Colorado School of Medicine, USA
Dina Demner-Fushman	National Library of Medicine, USA
Paul Thompson	University of Manchester, UK

Workshop Programme Committee

Jari Björne	University of Turku, Finland
Olivier Bodenreider	National Library of Medicine, USA
Wendy Chapman	UCSD, USA
Hongfang Liu	Mayo Clinic, USA
Naoaki Okazaki	Tohoku University, Japan
Sampo Pyysalo	University of Manchester, UK
Andrey Rzhetsky	University of Chicago, UK
Stefan Schulz	Medical University Graz, Austria
Lucy Vanderwende	Microsoft Research, USA
Karin Verspoor	NICTA, Australia
John Wilbur	NCBI, NLM, NIH, USA
Stephen Wu	Mayo Clinic, USA
Pierre Zweigenbaum	LIMSI, France

Introduction

Over the past decade, biomedical text mining has received a large amount of interest. Faced with the rapidly increasing volume of biomedical literature, domain experts have an ever-increasing need for tools that can help them locate isolate relevant nuggets of information from this deluge of information in a timely and efficient manner. The response to such issues by the natural language processing community can be clearly evidenced by the successful biomedical natural language processing workshops (BioNLP) that have been held over that past 10 years, in conjunction with ACL or NAACL meetings, to report the process in the field, as well as the founding of an ACL special interest group.

Biomedical text mining applications are reliant on high quality resources. These include databases and ontologies (e.g., Biothesaurus, UMLS Metathesaurus, MeSH and the Gene Ontology) and dictionaries/computational lexicons (e.g., the BioLexicon and the UMLS SPECIALIST lexicon). Recent years have also evidenced a large increase in the number of freely-available corpora (e.g., GENIA, GREC, AIMED, BioInfer, CRAFT, BioDRB) annotated with an expanding range of information types. These now include not only named entities and simple relations that hold between them, but also more complex event structures and coreference, as well as higher level information about how events are to be interpreted (e.g., facts, analyses, speculations, etc.) and discourse structure. Community shared tasks and challenges (e.g., JNLPBA, LL05, Biocreative I/II/III, BioNLP'09, BioNLP 2011, i2b2, etc.) also produce annotated corpora (on which the participating systems are trained and evaluated), in addition to steering research efforts to focus on open research problems. The development of high quality resources is very much relevant to META-NET (a Network of Excellence consisting of 54 research centres from 33 countries), that aims to stimulate a pan-European acceleration of research language technologies; this is dependent on the availability of appropriate resources.

The papers presented at the *3^rd^ Workshop on Building and Evaluating Resources for Biomedical Text Mining* exemplify the diversity of research that is currently taking place. Three papers concern resources for a relatively resource-poor language, i.e. Swedish. One of these describes a biomedical corpus derived from the Journal of the Swedish Medical Association (Kokkinakis), whilst the other two address de-identification of records in Swedish clinical corpora to remove Protected Health Information (PHI), using 2 different methods, i.e. pseudonymysation (Alfalahi et al.) and replacement of words with features (Dalianis and Boström). A third paper considering clinical corpora (Xia and Yetisgen-Yildiz) explains the challenges faced during annotation, and highlights the need for domain experts and detailed guidelines. In contrast, a further paper about annotation (Stubbs) proposes an annotation methodology for “light” annotation tasks for biomedical corpora, which do not require extensive training or exceptionally long annotation periods.

Three papers relate to biomedical relations or events. Kaewphan et al. describe the application of a literature-scale event extraction resource, EVEX, to NADP(H) metabolism regulation in *Escherichia coli*. The other two papers present new annotated corpora. Thomas et al. present two new corpora for protein-protein interactions and drug-drug interactions, which were automatically annotated, using distant supervision methods. Nawaz et al. describe the application of their multi-dimensional meta-knowledge annotation scheme to previously annotated biomedical events in a small collection of full papers, in order to enrich them with aspects of event interpretation such as negation, speculation, and knowledge source. The results are compared with a previous annotation effort for abstracts. The importance of recognising such interpretative information in biomedical texts is reinforced in the paper by Makkaoui et al., which evaluates a system for annotating speculative sentences on the BioScope corpus.

The remaining two papers in this volume concern named entity annotations. Neves et al. present a corpus for stem cell research, which is annotated with different types of entities relevant to this subdomain. Preliminary results of automatic recognition of these entities are also presented. Mihăilă et al. examine the distribution of named entity types across 20 different biomedical subdomains. The degree of difference or similarity between different subdomains can be an important consideration when adapting automated tools from one subdomain to another.

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, Jun'ichi Tsujii, for his contribution.

Sophia Ananiadou, Kevin Cohen, Dina Demner-Fushman and Paul Thompson

Session 1

Saturday 26 May, 11:00 – 12:15
Chairperson: Paul Thompson

Analysing Entity Type Variation across Biomedical Subdomains

Claudiu Mihăilă, Riza Theresa Batista-Navarro and Sophia Ananiadou

Previous studies have shown that various biomedical subdomains have lexical, syntactic, semantic and discourse structure variations. It is essential to recognise such differences to understand that biomedical natural language processing tools, such as named entity recognisers, that work well on some subdomains may not work as well on others. In this paper, we investigate the pairwise similarity (or dissimilarity) amongst twenty selected biomedical subdomains, at the level of named entity types. We evaluate the contribution of these types in the classification task by computing the chi-squared statistic over their distributions. We then build a binary classifier for each possible pair of subdomains, the results of which indicate the subdomains that are highly different or similar to others. The findings can be of potential use to those building or using named entity recognisers in determining which types of named entities need to be taken into consideration or in adapting already existing tools.

Integrating Large-Scale Text Mining and Co-Expression Networks: Targeting NADP(H) Metabolism in E. coli with Event Extraction

Suwisa Kaewphan, Sanna Kreula, Sofie Van Landeghem, Yves Van de Peer, Patrik R. Jones and Filip Ginter

We present an application of EVEX, a literature-scale event extraction resource, in the concrete biological use case of NADP(H) metabolism regulation in *Escherichia coli.* We make extensive use of the EVEX event generalization based on gene family definitions in Ensembl Genomes, to extract cross-species candidate regulators. We manually evaluate the resulting network so as to only preserve correct events and facilitate its integration with microarray-based co-expression data. When analysing the combined network obtained from text mining and co-expression, we identify 41 candidate genes involved in triangular patterns involving both subnetworks. Several of these candidates are of particular interest, and we discuss their biological relevance further. This study is the first to present a real-world evaluation of the EVEX resource in particular and literature-scale application of the systems emerging from the BioNLP Shared Task series in general. We summarize the lessons learned from this use case in order to focus future development of EVEX and similar literature-scale resources.

Annotating and Evaluating Text for Stem Cell Research

Mariana Neves, Alexander Damaschun, Andreas Kurtz and Ulf Leser

The regeneration of vital organs and tissues remains one of the biggest medical challenges. However, the use of embryonic stem cells and induced pluripotent stem cells allows novel replacement strategies. The CellFinder project aims to create a stem cell data repository by linking information from existing public databases and by performing text mining on the research literature. We present the first version of our corpus which is composed of 10 full text documents containing more than 2,100 sentences, 65,000 tokens and 5,200 annotations for entities. The corpus has been annotated with six types of entities (anatomical parts, cell components, cell lines, cell types, genes/protein and species) with an overall inter-annotator agreement around 80%. Preliminary results using baseline methods based on freely available terminologies and systems have returned a recall which ranges from 48% to 90% for the extraction of the named entities. The high distribution of entities which are representative of the stem cell research, specially cell types, makes our corpus a valuable resource for the stem cell domain.

Session 2

Saturday 26 May, 14:00 – 15:15
Chairperson: Kevin Cohen

Meta-Knowledge Annotation at the Event Level: Comparison between Abstracts and Full Papers

Raheel Nawaz, Paul Thompson and Sophia Ananiadou

Biomedical literature contains rich information about events of biological relevance. Event corpora, containing classified, structured representations of important facts and findings contained within text, provide an important resource for the training of domain-specific information extraction (IE) systems. Such corpora pay little attention to the interpretation of events, e.g., whether an event describes a fact or an analysis of results, whether there is any speculation surrounding the event, etc. These types of information are collectively referred to as *meta-knowledge*. As previous work, an annotation scheme to enrich event corpora with meta-knowledge was designed to facilitate the training of more sophisticated IE systems, and was applied to the complete GENIA Event corpus of biomedical abstracts. In this paper, we describe a case study in which four full papers annotated with GENIA events have been manually enriched with meta-knowledge annotation. We analyse the annotation results, and compare them with the previously annotated abstracts.

Clinical Corpus Annotation: Challenges and Strategies

Fei Xia and Meliha Yetisgen-Yildiz

Annotation is an important task for Natural Language Processing (NLP), and the traditional annotation schema, including writing detailed guidelines and training annotators, has proved to work well in many previous annotation projects. However, making medical judgment on clinical data requires medical expertise and annotation can only be done by experts. Recently, we created three corpora for our clinical NLP studies: one marks critical recommendations in radiology reports, and the other two indicate whether a patient has pneumonia based on chest X-ray reports or ICU reports. All the annotations were done by medical experts. In this paper, we discuss various challenges we have encountered when dealing with expert annotation, and lay out some lessons we have learned from the annotation tasks. Our experiments show that medical training alone is not sufficient for achieving high inter-annotator agreement, and NLP researchers should get involved in the annotation process as early as possible despite their lack of medical training.

The Journal of the Swedish Medical Association - a Corpus Resource for Biomedical Text Mining in Swedish

Dimitrios Kokkinakis

Biomedical text mining applications are largely dependent on high quality knowledge resources. Traditionally, these resources include lexical databases, terminologies, nomenclatures and ontologies and, during the last decade, also corpora of various sizes, variety and diversity. Some of these corpora are annotated with an expanding range of information types and metadata while others become available with a minimal set of annotations. It is also of great importance that biomedical corpora for lesser-spoken languages also get developed. This is required in order to support and facilitate implementation of practical applications for such languages and to stimulate the development of language technology research and innovation infrastructures in the domain. This paper provides a description of a Swedish biomedical corpus based on the electronic editions of the *Journal of the Swedish Medical Association* "Läkartidningen" of the years 1996-2010. The corpus consists of a variety of documents that can be related to different medical domains, developed as a response to the increasing needs for large and reliable medical information for Swedish biomedical Natural Language Processing (NLP). The corpus has been structurally annotated with a minimal set of meta information and automatically indexed with the *Swedish Systematized Nomenclature of Medicine -- Clinical Terms* (SNOMED CT).

Session 3 – Poster Session

Saturday 26 May, 15:15 – 16:30
Chairperson: Paul Thompson

Releasing a Swedish Clinical Corpus after Removing All Words - De-Identification Experiments with Conditional Random Fields and Random Forests

Hercules Dalianis and Henrik Boström

Patient records contain valuable information in the form of both structured data and free text; however this information is sensitive since it can reveal the identity of patients. In order to allow new methods and techniques to be developed and evaluated on real world clinical data without revealing such sensitive information, researchers could be given access to de-identified records without protected health information (PHI), such as names, telephone numbers, and so on. One approach to minimizing the risk of revealing PHI when releasing text corpora from such records is to include only features of the words instead of the words themselves. Such features may include parts of speech, word length, and so on from which the sensitive information cannot be derived. In order to investigate what performance losses can be expected when replacing specific words with features, an experiment with two state-of-the-art machine learning methods, conditional random fields and random forests, is presented, comparing their ability to support de-identification, using the Stockholm EPR PHI corpus as a benchmark test. The results indicate severe performance losses when the actual words are removed, leading to the conclusion that the chosen features are not sufficient for the suggested approach to be viable.

Pseudonymisation of Personal Names and other PHIs in an Annotated Clinical Swedish Corpus

Alyaa Alfalahi, Sara Brissman and Hercules Dalianis

Today a large number of patient records are produced and these records contain valuable information, often in free text, about the medical treatment of patients. Since these records contain information that can reveal the identity of patients, known as protected health information (PHI), the records cannot easily be made available for the research community. In this research we have used a PHI annotated clinical corpora, written in Swedish, that we have pseudonymised. Pseudonymisation means to replace the sensitive information with fictive information for example real personal names are replaced with fictive personal names based on the gender of the real names and family relations. We have evaluated our results and our five respondents of who three were clinicians found that the clinical text looks real and is readable. We have also added pseudonymisation for telephone numbers, locations, health care units, dates and ages. In this paper we also present the entire de-identification and pseudonymisation process of a sample clinical text.

Evaluation and Performance Improvement of the BioExcom System for the Automatic Detection of Speculation in Biomedical Texts

Olfa Makkaoui, Julien Desclés and Jean-Pierre Desclés

The BioExcom system aims to automatically annotate speculative sentences in biomedical texts and to categorize them into "*new*" and "*prior*" speculations. This work highlights a more restrictive way to consider speculations as a source of knowledge for biologists who are also interested in finding hypotheses in the biomedical literature. The system is based on the Contextual Exploration processing (hierarchical research of linguistic surface markers with the EXCOM computational platform). The BioExcom evaluation is realized on the BioScope corpus by manually comparing the BioExcom automatic annotations and the BioScope manual annotations. Theanalysis of diverging annotations was a starting point to build a new version of the system (BioExcom_2) that results from the performance improvement of the initial system BioExcom. A corpus *BioSpe* for the annotation of speculative sentences is established. This latter was annotated according the BioExcom characterization of speculation and can be used by machine learning systems. A user interface for the automatic annotation of speculative sentences is made available on line.

Session 4

Saturday 26 May, 16:30 – 17:20
Chairperson: Sophia Ananiadou

Weakly Labeled Corpora as Silver Standard for Drug-Drug and Protein-Protein Interaction

Philippe Thomas, Tamara Bobić, Martin Hofmann-Apitius, Ulf Leser and Roman Klinger

Relation extraction is frequently and successfully addressed by machine learning methods. The downside of this approach is the need for annotated training data, typically generated in tedious manual, cost intensive work. Distantly supervised approaches make use of weakly annotated data, which can be derived automatically. Recent work in the biomedical domain has applied distant supervision for protein-protein interaction (PPI) with reasonable results, by employing the IntAct database. Training from distantly labeled corpora is more challenging than from manually curated ones, as such data is inherently noisy. With this paper, we make two corpora publicly available to the community to allow for comparison of different methods that deal with the noise in a uniform setting. The first corpus is addressing protein-protein interaction (PPI), based on named entity recognition and the use of IntAct and KUPS databases, the second is concerned with drug-drug interaction (DDI), making use of the database DrugBank. Both corpora are in addition labeled with 5 state-of-the-art classifiers trained on annotated data, to allow for development of filter methods. Furthermore, we present in short our approach and results for distant supervision on these corpora as a strong baseline for future research.

Developing Specifications for Light Annotation Tasks in the Biomedical Domain

Amber Stubbs

Biomedical texts pose an interesting challenge in natural language processing tasks. While the information contained in them is important to people of all backgrounds, often they are stylistically complex with specialized vocabularies, and require advanced degrees or other special training to interpret correctly. Because of this, researchers in Natural Language Processing are often at a disadvantage when it comes to extracting task-specific information from these texts: the experts who are best able to understand them may not have the time or interest in completing complicated and time-consuming annotations for use in corpus analysis and machine learning. This paper proposes a methodology for creating light annotation tasks for biomedical corpora that can be used to create useful annotations without requiring extensive training or exceptionally long annotation periods. The utility of the proposed methodology is examined in light of existing annotation projects, as well as through the lens of a case study using hospital discharge summaries for patient selection based on eligibility criteria.

Language Engineering for Online Reputation Management

26 May 2012

ABSTRACTS

Editors:

Adolfo Corujo, Julio Gonzalo, Edgar Meij, Maarten de Rijke, Irina Chugur

Workshop Programme

9:00 - 9:45 – Introduction to the NLP & ORM Challenge
Julio Gonzalo (UNED), *The RepLab Initiative: An Evaluation Campaign for Online Reputation Management*

Hugo Zaragoza (WebSays), *Online Reputation Management: Business Requirements and Scientific Challenges*

Miguel Lucas (Acteo), *Online Reputation Management: Analysis of Existing Commercial Tools*

9:45 - 10:30 – Position papers
Alexandra Balahur (JRC), *The Challenge of Processing Opinions in Online Contents in the Social Web Era*

Patrick Brennan (Juola & Associates), *Tagging Commentary with Demographic Data*

Fredrik Olsson, Jussi Karlgren, Magnus Sahlgren, Fredrik Espinoza, Ola Hamfors, (Gavagai), *Technical Requirements For Knowledge Representation For Reputation Mining On A Realistic Scale*

10:30 - 11:00 – Coffee Break

11:00 - 11:45 – Technical Papers
Chandra Mohan Dasari, Dipankar Das, Sivaji Bandyopadhyay (Jadavpur University), *Topic Identification from Blog Documents: Roles of Bigram, Named Entity and Sentiment*

Yue Dai, Ernest Aredarenko, Tuomo Kakkonen, Ding Liao (University of Eastern Finland), *Towards SoMEST – Combining Social Media Monitoring with Event Extraction and Timeline Analysis*

Damiano Spina (UNED), Edgar Meij, Andrei Oghina, Minh Thuong Bui, Mathias Breuss, Maarten de Rijke (University of Amsterdam), *A Corpus for Entity Profiling in Microblog Posts*

11:45 – 13:00 – Roadmap Discussion & Wrap-up
Jordi Atserias (Yahoo! Research Barcelona)
Adolfo Corujo (Llorente & Cuenca)
Julio Gonzalo (UNED)
Miguel Lucas (Acteo)
Edgar Meij (University of Amsterdam)
Maarten de Rijke (University of Amsterdam)
Hugo Zaragoza (WebSays) plus all workshop participants

Workshop Organizers

Adolfo Corujo	Llorente & Cuenca, Spain
Julio Gonzalo	UNED, Spain
Edgar Meij	University of Amsterdam, The Netherlands
Maarten de Rijke	University of Amsterdam, The Netherlands

Workshop Programme Committee

Eugene Agichtein	Emory University, USA
Alexandra Balahur	JRC, Italy
Krisztian Balog	NTNU, Norway
Raymond Franz	Trendlight, The Netherlands
Donna Harman	NIST, USA
Eduard Hovy	ISI/USC, USA
Radu Jurca	Google, Switzerland
Jussi Karlgren	Gavagai/SICS, Sweden
Mounia Lalmas	Yahoo! Research, Spain
Jochen Leidner	Thomson Reuters, Switzerland
Bing Liu	U. Illinois at Chicago, USA
Alessandro Moschitti	U. Trento, Italy
Miles Osborne	U. Edinburgh, UK
Hans Uszkoreit	U. Saarbrucken, Germany
James Shanahan	Boston U., USA
Belle Tseng	Yahoo!, USA
Julio Villena	Daedalus/U. Carlos III, Spain

Preface

This volume collects technical and position papers for the LREC Workshop on Language Engineering for Online Reputation Management held in Istanbul on May 26, 2012.

Online Reputation Management deals with the image that online media project about individuals and organizations. The growing relevance of social media and the speed at which facts and opinions travel in microblogging networks make online reputation an essential part of a company's public relations.

While traditional reputation analysis was based mostly on manual analysis (clipping from media, surveys, etc.), the key value from online media comes from the ability of processing, understanding and aggregating potentially huge streams of facts and opinions about a company or individual. Information to be mined includes answers to questions such as: What is the general state of opinion about a company/individual in online media? What are its perceived strengths and weaknesses, as compared to its peers/competitors? How is the company positioned with respect to its strategic market? Can incoming threats to its reputation be detected early enough to be neutralized before they effectively affect reputation?

In this context, Natural Language Processing plays a key, enabling role, and we are already witnessing an unprecedented demand for text mining software in this area. Note that, while the area of opinion mining has made significant advances in the last few years, most tangible progress has been focused on products. However, mining and understanding opinions about companies and individuals is, in general, a much harder and less understood problem.

The aim of the workshop was to bring together the Language Engineering community (including researchers and developers) with representatives from the Online Reputation Management industry, a fast-growing sector which poses challenging demands to text mining technologies. The goal was to establish a five-year roadmap on the topic, focusing on what language technologies are required to get there in terms of resources, algorithms and applications. The workshop is tightly connected to RepLab, an evaluation initiative for Online Reputation Management Systems which has its first edition as a CLEF 2012 lab, in September 2012. The outcome of the workshop is intended to serve as direct input to establish the research priorities of RepLab.

With this purpose in mind, the workshop included both research papers and position statements from industry and academia. Besides paper presentations, the agenda of the workshop includes a session introducing the problem from a dual business and academic perspective, and a discussion session aimed at establishing a roadmap for the topic. The workshop is partially supported by the EU project Limosine (under project number 288024, call FP7-ICT-2011-7).

Position Papers

9:45 - 10:30

The Challenge of Processing Opinions Expressed in Online Contents in the Social Web Era

Alexandra Balahur

Abstract

In the new Social Web era, the globalization of markets combined with the fact that people can freely express their opinion on any product or company on forums, blogs or e-commerce sites led to a change in the companies' marketing strategies, in the rise of awareness for client needs and complaints, and a special attention for brand trust and reputation. Specialists in market analysis, but also IT fields such as Natural Language Processing (NLP), demonstrated that in the context of the newly created opinion phenomena, decisions for economic action are not only given by factual information, but are highly affected by rumors and negative opinions. In this context, analyzing online reputation and being able to understand the mechanisms through which opinions are spread and the extent and manner in which they influence the business, social and political spheres become necessary endeavors. The problem in this context is much more difficult to solve, as entities, as opposed to products, are related to different events and topics and there is no fixed set of "attributes" that are commented on by persons expressing opinions on these entities. Additionally, answering opinion questions is an issue that is far from being trivial. This paper describes the challenges related to mining opinions for reputation management in the Social Web context.

Technical Requirements for Knowledge Representation for Attitude Mining on a Realistic Scale

Fredrik Olsson, Jussi Karlgren, Magnus Sahlgren, Fredrik Espinoza, Ola Hamfors

Abstract

To be useful, a reputation mining system must cover a broad range of weakly, vaguely, and implicitly expressed human sentiments and cannot in the absence of prior knowledge rely on sampling the data stream of human-generated text. To achieve coverage, a reputation mining system must be sensitive to variation and change in the signal. These requirements pose a challenge that are an instance of more general semantic processing – this paper presents some design requirements used to design and implement a semantic layer for a processing stack for human-generated information.

Uses of Computational Stylometry to Determine Demographics for Online Reputation Management

Patrick Brennan

Abstract

Computational stylometry can be used as a tool to gather better demographic data for the purposes of online reputation management. Computational stylometry is the study of linguistically style; in this case, applied to blog posts and comments on web sites. These sorts of messages are generally both anonymous and honest appraisals of products and services, so being able to gather more data about who these comments represent will provide businesses with a better idea of how they are doing with these demographic groups. Finally, we illustrate a product currently in the works that

will tag comments with demographic tags through the use of computational stylometry.

Technical Papers

11:00 - 11:45

Topic Identification from Blog Documents: Roles of Bigram, Named Entity and Sentiment

Chandra Mohan Dasari, Dipankar Das, Sivaji Bandyopadhyay

Abstract

The rapid growth of blog documents in Web 2.0 and categorizing search applications based on topics motivates us to develop a system that identifies topic names of the blog documents using Bigrams, Named Entity (NE) and Sentiment features. We also associate the sentiment scores to the blog documents using the SentiWordNet. The individual module based on Bigrams, NE and Sentiment produces the topic bag for each blog document containing probable topic names of that blog. The probable topic names were evaluated manually based on top-n (n = 5, 10 and 20) ranking mechanism. Though the combined module of Bigram and Sentiment performs better than the combined module of Bigram and NE, the combination of all the three modules produces satisfactory results on evaluating 125 topic names with respect to 25 test documents on 5 different topics. The evaluation achieves the maximum accuracies of 60.0%, 72.0% and 84.0% for the combined module of Bigram and Sentiment and 76.0%, 86.0% and 92.0% for the combined module of Bigram and Named Entity with respect to top-5, top-10 and top-20 ranking mechanisms, respectively.

Towards SoMEST – Combining Social Media Monitoring with Event Extraction and Timeline Analysis

Yue Dai, Ernest Aredarenko, Tuomo Kakkonen, Ding Liao

Abstract

We report on the development of a social media monitoring tool based on the novel Social Media Event Sentiment Timeline (SoMEST) model. The novelty of our model is that it combines opinion mining techniques with a timeline-based event analysis method and an information and event extraction tool. While Event Timeline Analysis (ETA) is an existing method utilized in analyzing the external environment of businesses, the SoMEST model and the BEECON (Business Events Extractor Component based on Ontology) tool as well as the OMS (Opinion Miner for SoMEST) component we report on are developed by the authors of the current paper.

A Corpus for Entity Profiling in Microblog Posts

Damiano Spina, Edgar Meij, Andrei Oghina, Minh Thuong Bui, Mathias Breuss, Maarten de Rijke

Abstract

Microblogs have become an invaluable source of information for the purpose of online reputation management. Streams of microblogs are of great value because of their direct and real-time nature.

An emerging problem is to identify not only microblog posts (such as tweets) that are relevant for a given entity, but also the specific aspects that people discuss. Determining such aspects can be non-trivial because of creative language usage, the highly contextualized and informal nature of microblog posts, and the limited length of this form of communication. In this paper we present two manually annotated corpora to evaluate the task of identifying aspects on Twitter, both of them based upon the WePS-3 ORM task dataset and made available online. The first is created using a pooling methodology, for which we have implemented various methods for automatically extracting aspects from tweets that are relevant for an entity. Human assessors have labeled each of the candidates as being relevant. The second corpus is more fine-grained and contains opinion targets. Here, annotators consider individual tweets related to an entity and manually identify whether the tweet is opinionated and, if so, which part of the tweet is subjective and what the target of the sentiment is, if any.

The 5th Workshop on Building and Using Comparable Corpora

Special Theme: “Language Resources for
Machine Translation in Less-Resourced Languages and Domains”

LREC2012 Workshop
26 May 2012
Istanbul, Turkey

ABSTRACTS

Editors:

Reinhard Rapp, Marko Tadić, Serge Sharoff, Pierre Zweigenbaum

Workshop Programme

09:00 – 09:10 **Opening**

Oral Presentations 1: Multilinguality (Chair: Pierre Zweigenbaum)

09:10 – 09:30 Philipp Petrenz, Bonnie Webber: *Robust Cross-Lingual Genre Classification through Comparable Corpora*

09:30 – 09:50 Qian Yu, François Yvon, Aurélien Max: *Revisiting sentence alignment algorithms for alignment visualization and evaluation*

Invited Projects Session (Chair: Serge Sharoff)

09:50 – 10:10 Inguna Skadiņa: *Analysis and Evaluation of Comparable Corpora for Under-Resourced Areas of Machine Translation* (ACCURAT, http://www.accurat-project.eu)

10:10 – 10:30 Andrejs Vasiļjevs: *LetsMT! – Platform to Drive Development and Application of Statistical Machine Translation* (LetsMT!, http://www.letsmt.eu)

10:30 – 11:00 **Coffee Break**

Invited Project Session (Contd.)

11:00 – 11:20 Núria Bel, Vassilis Papavasiliou, Prokopis Prokopidis, Antonio Toral, Victoria Arranz: *Mining and Exploiting Domain-Specific Corpora in the PANACEA Platform* (PANACEA, http://panacea-lr.eu)

11:20 – 11:40 Adam Kilgarriff, George Tambouratzis: *The PRESEMT Project* (PRESEMT, http://www.presemt.eu)

11:40 – 12:00 Béatrice Daille: *Building Bilingual Terminologies from Comparable Corpora: The TTC TermSuite* (TTC, http://www.ttc-project.eu)

12:00 – 12:30 **Panel Discussion with Invited Speakers**

12:30 – 14:00 **Lunch Break**

Oral Presentations 2: Building Comparable Corpora (Chair: Reinhard Rapp)

14:00 – 14:20 Aimée Lahaussois, Séverine Guillaume: *A viewing and processing tool for the analysis of a comparable corpus of Kiranti mythology*

14:20 – 14:40 Nancy Ide: *MultiMASC: An Open Linguistic Infrastructure for Language Research*

Booster Session for Posters (Chair: Marko Tadić)

14:40 – 14:45 Elena Irimia: *Experimenting with Extracting Lexical Dictionaries from Comparable Corpora for: English-Romanian language pair*

14:45 – 14:50 Iustina Ilisei, Diana Inkpen, Gloria Corpas, Ruslan Mitkov: *Romanian Translational Corpora: Building Comparable Corpora for Translation Studies*

14:50 – 14:55 Angelina Ivanova: *Evaluation of a Bilingual Dictionary Extracted from Wikipedia*

14:55 – 15:00 Quoc Hung-Ngo, Werner Winiwarter: *A Visualizing Annotation Tool for Semi-Automatical Building a Bilingual Corpus*

15:00 – 15:05 Lene Offersgaard, Dorte Haltrup Hansen: *SMT systems for less-resourced languages based on domain-specific data*

15:05 – 15:10 Magdalena Plamada, Martin Volk: *Towards a Wikipedia-extracted Alpine Corpus*

15:10 – 15:15 Sanja Štajner, Ruslan Mitkov: *Using Comparable Corpora to Track Diachronic and Synchronic Changes in Lexical Density and Lexical Richness*

15:15 – 15:20 Dan Ştefănescu: *Mining for Term Translations in Comparable Corpora*

15:20 – 15:25 George Tambouratzis, Michalis Troullinos, Sokratis Sofianopoulos, Marina Vassiliou: *Accurate phrase alignment in a bilingual corpus for EBMT systems*

15:25 – 15:30 Kateřina Veselovská, Ngụy Giang Linh, Michal Novák: *Using Czech-English Parallel Corpora in Automatic Identification of* It

15:30 – 15:35 Manuela Yapomo, Gloria Corpas, Ruslan Mitkov: *CLIR- and Ontology-Based Approach for Bilingual Extraction of Comparable Documents*

15:35 – 16:30 **Poster Session and Coffee Break** (coffee from 16:00 – 16:30)

Oral Presentations 3: Lexicon Extraction and Corpus Analysis (Chair: Andrejs Vasiļjevs)

16:30 – 16:50 Amir Hazem, Emmanuel Morin: *ICA for Bilingual Lexicon Extraction from Comparable Corpora*

16:50 – 17:10 Hiroyuki Kaji, Takashi Tsunakawa, Yoshihoro Komatsubara: *Improving Compositional Translation with Comparable Corpora*

17:10 – 17:30 Nikola Ljubešić, Špela Vintar, Darja Fišer: *Multi-word term extraction from comparable corpora by combining contextual and constituent clues*

17:30 – 17:50 Robert Remus, Mathias Bank: *Textual Characteristics of Different-sized Corpora*

17:50 – 18:00 **Wrapup discussion and end of the workshop**

Workshop Organizing Committee

Reinhard Rapp	University of Leeds and University of Mainz
Marko Tadić	University of Zagreb, Faculty of Humanities and Social Sciences
Serge Sharoff	University of Leeds
Pierre Zweigenbaum	LIMSI-CNRS and ERTIM-INALCO, Orsay
Andrejs Vasiļjevs	Tilde, Riga

Workshop Programme Committee

Srinivas Bangalore	AT&T Labs, USA
Caroline Barrière	National Research Council Canada
Chris Biemann	Microsoft / Powerset, San Francisco, USA
Lynne Bowker	University of Ottawa, Canada
Hervé Déjean	Xerox Research Centre Europe, Grenoble, France
Andreas Eisele	DFKI, Saarbrücken, Germany
Rob Gaizauskas	University of Sheffield, UK
Éric Gaussier	Université Joseph Fourier, Grenoble, France
Nikos Glaros	ILSP, Athens, Greece
Gregory Grefenstette	Exalead/Dassault Systemes, Paris, France
Silvia Hansen-Schirra	University of Mainz, Germany
Kyo Kageura	University of Tokyo, Japan
Adam Kilgarriff	Lexical Computing Ltd, UK
Natalie Kübler	Université Paris Diderot, France
Philippe Langlais	Université de Montréal, Canada
Tony McEnery	Lancaster University, UK
Emmanuel Morin	Université de Nantes, France
Dragos Stefan Munteanu	Language Weaver Inc., USA
Lene Offersgaard	University of Copenhagen, Denmark
Reinhard Rapp	Universities of Mainz, Germany, and Leeds, UK
Sujith Ravi	Yahoo! Research, Santa Clara, CA, USA
Serge Sharoff	University of Leeds, UK
Michel Simard	National Research Council Canada
Inguna Skadiņa	Tilde, Riga, Latvia
Monique Slodzian	INALCO, Paris, France
Benjamin Tsou	The Hong Kong Institute of Education, China
Dan Tufis	Romanian Academy, Bucharest, Romania
Justin Washtell	University of Leeds, UK
Michael Zock	LIF, CNRS Marseille, France
Pierre Zweigenbaum	LIMSI-CNRS and ERTIM-INALCO, Orsay, France

Invited Speakers

Núria Bel	University Pompeu Fabra, Barcelona, Spain
Béatrice Daille	University of Nantes, France
Adam Kilgarriff	Lexical Computing Ltd., UK
Inguna Skadiņa	Tilde, Riga, Latvia
Andrejs Vasiļjevs	Tilde, Riga, Latvia

Preface

Following the four previous editions of the Workshop on Building and Using Comparable Corpora which took place at LREC 2008 in Marrakech, at ACL-IJCNLP 2009 in Singapore, at LREC 2010 in Malta, and at ACL-HLT 2011 in Portland, this year the workshop was co-located with LREC 2012 in Istanbul.

Although papers on all topics related to comparable corpora were welcome at the workshop, this year's special theme was "Language Resources for Machine Translation in Less-Resourced Languages and Domains". This theme was chosen with the aim of finding ways to overcome the shortage of parallel resources when building machine translation systems for less-resourced languages and domains. Lack of sufficient language resources for many language pairs and domains is currently one of the major obstacles in the further advancement of machine translation. Possible solutions include the identification of parallel segments within comparable corpora or reaching out for parallel data that is 'hidden' in users' repositories.

To highlight the increasing interest in comparable corpora and the success of the field, representatives from five international research projects were invited to present the important role of work on comparable corpora within a special session. These projects were ACCURAT (http://www.accurat-project.eu/), LetsMT! (https://www.letsmt.eu/), PANACEA (http://panacea-lr.eu/), PRESEMT (http://www.presemt.eu/), and TTC (http://www.ttc-project.eu/).

We would like to thank all people and institutions who helped in making this workshop a success. This year the workshop has been formally endorsed by ACL SIGWAC (Special Interest Group on Web as Corpus), FLaReNet (Fostering Language Resources Network), and META-NET (Multilingual Europe Technology Alliance). Our special thanks go to the representatives of the above mentioned projects for accepting our invitations, to the members of the program committee who did an excellent job in reviewing the submitted papers under strict time constraints, and to the organizers of the hosting conference. Last but not least we would like to thank our authors and the participants of the workshop.

Reinhard Rapp
Marko Tadić
Serge Sharoff
Pierre Zweigenbaum

Oral presentations 1: Multilinguality

Saturday 26 May 2012, 09:10 – 9:50
Chairperson: Pierre Zweigenbaum

Robust Cross-Lingual Genre Classification through Comparable Corpora

Philipp Petrenz, Bonnie Webber

Classification of texts by genre can benefit applications in Natural Language Processing and Information Retrieval. However, a mono-lingual approach requires large amounts of labeled texts in the target language. Work reported here shows that the benefits of genre classification can be extended to other languages through cross-lingual methods. Comparable corpora – here taken to be collections of texts from the same set of genres but written in different languages – are exploited to train classification models on multi-lingual text collections. The resulting genre classifiers are shown to be robust and high-performing when compared to mono-lingual training sets. The work also shows that comparable corpora can be used to identify features that are indicative of genre in various languages. These features can be considered stable genre predictors across a set of languages. Our experiments show that selecting stable features yields significant accuracy gains over the full feature set, and that a small amount of features can suffice to reliably distinguish between different genres.

Revisiting sentence alignment algorithms for alignment visualization and evaluation

Qian Yu, François Yvon, Aurélien Max

In this paper, we revisit the well-known problem of sentence alignment, in a context where the entire bitext has to be aligned and where alignment confidence measures have to be computed. Following much recent work, we study here a multi-pass approach: we first compute sure alignments that are used to train a discriminative model; then we use this model to fill in the gaps between sure links. Experimental results on several corpora show the effectiveness of this method as compared to alternative, state-of-the-art, proposals.

Invited projects session

Saturday 26 May 2012, 09:50 – 10:30
Chairperson: Pierre Zweigenbaum

Analysis and Evaluation of Comparable Corpora for Under-Resourced Areas of Machine Translation

Inguna Skadiņa

This abstract presents the FP7 project ACCURAT that aims to research methods and create tools that find, measure, and use bi/multilingual comparable corpora to improve the quality of machine translation for under-resourced languages and narrow domains. Work on corpora collection, assessment of the comparability of documents pairs in collected corpora, extraction of parallel data for the machine translation (MT) task, and application to the MT task is presented.
ACCURAT, http://www.accurat-project.eu

LetsMT! – Platform to Drive Development and Application of Statistical Machine Translation

Andrejs Vasiļjevs

This paper presents ICT-PSP project LetsMT! which develops a user-driven machine translation "factory on the cloud". Current mass-market and online MT systems are of general nature, system adaptation for specific needs is prohibitively expensive service not affordable to smaller companies or public institutions. To exploit the huge potential of open statistical machine translation (SMT) technologies LetsMT! has created an innovative online collaborative platform for data sharing and MT building.
LetsMT!, http://www.letsmt.eu

Invited projects session (continued)

Saturday 26 May 2012, 11:00 – 12:00
Chairperson: Serge Sharoff

Mining and Exploiting Domain-Specific Corpora in the PANACEA Platform

Núria Bel, Vassilis Papavasiliou, Prokopis Prokopidis, Antonio Toral, Victoria Arranz

The objective of the PANACEA ICT-2007.2.2 EU project is to build a platform that automates the stages involved in the acquisition, production, updating and maintenance of the large language resources required by, among others, MT systems. The development of a Corpus Acquisition Component (CAC) for extracting monolingual and bilingual data from the web is one of the most innovative building blocks of PANACEA. The CAC, which is the first stage in the PANACEA pipeline for building Language Resources, adopts an efficient and distributed methodology to crawl for web documents with rich textual content in specific languages and predefined domains. The CAC includes modules that can acquire parallel data from sites with in-domain content available in more than one language. In order to extrinsically evaluate the CAC methodology, we have conducted several experiments that used crawled parallel corpora for the identification and extraction of parallel sentences using sentence alignment. The corpora were then successfully used for domain adaptation of Machine Translation Systems.
PANACEA, http://panacea-lr.eu

The PRESEMT Project

Adam Kilgarriff, George Tambouratzis

Within the PRESEMT project, we have explored a hybrid approach to machine translation in which a small parallel corpus is used to learn mapping rules between grammatical constructions in the two languages, and large target-language corpora are used for refining translations. We have also taken forward methods for 'corpus measurement', including an implemented framework for measuring the distance between any two corpora of the same language. We briefly describe developments in both these areas.
PRESEMT, http://www.presemt.eu

Building bilingual terminologies from comparable corpora: The TTC TermSuite

Béatrice Daille

In this paper, we exploit domain-specific comparable corpora to build bilingual terminologies. We present the monolingual term extraction and the bilingual alignment that will allow us to identify and translate high specialised terminology. We stress the huge importance of taking into account both simple and complex terms in a multilingual environment. Such linguistic diversity implies to combine several methods to perfect accurately both monolingual and bilingual terminology extraction tasks. The methods are implemented in TTC TermSuite based on a UIMA framework.
TTC, http://www.ttc-project.eu

Panel discussion with invited speakers

Saturday 26 May 2012, 12:00 – 12:30
Chairperson: Serge Sharoff

Oral Presentations 2: Building Comparable Corpora

Saturday 26 May 2012, 14:00 – 14:40
Chairperson: Reinhard Rapp

A viewing and processing tool for the analysis of a comparable corpus of Kiranti mythology

Aimée Lahaussois, Séverine Guillaume

This presentation describes a trilingual corpus of three endangered languages of the Kiranti group (Tibeto-Burman family) from Eastern Nepal. The languages, which are exclusively oral, share a rich mythology, and it is thus possible to build a corpus of the same native narrative material in the three languages. The segments of similar semantic content are tagged with a "similarity" label to identify correspondences among the three language versions of the story. An interface has been developed to allow these similarities to be viewed together, in order to allow make possible comparison of the different lexical and morphosyntactic features of each language. A concordancer makes it possible to see the various occurrences of words or glosses, and to further compare and contrast the languages.

MultiMASC: An Open Linguistic Infrastructure for Language Research

Nancy Ide

This paper describes MultiMASC, which builds upon the Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008; Ide et al., 2010) project, a community-based collaborative effort to create, annotate, and validate linguistic data and annotations on a broad-genre open language data. MultiMASC will extend MASC to include comparable corpora in other languages that not only represent the same genres and styles, but also include similar types and number of annotations represented in a common format. Like MASC, MultiMASC will contain only completely open data, and will rely on a collaborative community-based effort for its development. We describe the possible ways in which additional corpora for MultiMASC can be collected and annotated and consider the dimensions along which “comparability” for MultiMASC corpora can be determined. Because it is unlikely that all language-specific MultiMASC corpora can be comparable along every dimension, we also outline the measures that can be used to gauge comparability for a number of different criteria.

Booster Session for Posters

Saturday 26 May 2012, 14:40 – 15:35
Chairperson: Marko Tadić

Experimenting with Extracting Lexical Dictionaries from Comparable Corpora for: English-Romanian language pair

Elena Irimia

The paper describes a tool developed in the context of the ACCURAT project (Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation). The purpose of the tool is to extract bilingual lexical dictionaries (word-to-word) from comparable corpora which do not have to be aligned at any level (document, paragraph, etc.) The method implemented in this tool is introduced by (Rapp, 1999). The application basically counts word co-occurrences between unknown words in the comparable corpora and known words from a Moses extracted general domain translation table (the base lexicon). We adapted the algorithm to work with polysemous entries in the translation table, a very frequent situation which is not treated in the standard approach. We introduced other heuristics, like 1. filtration of the context vectors according to a log likelihood threshold, 2. lists of verbs (specific to each language) that can be main verbs but also auxiliary or modal verbs; 3) a cognate heuristic based on the Levenshtein Distance. The implementation can also run in multithreading mode, if the user’s machine has the capacity to enable parallel execution.

Romanian Translational Corpora: Building Comparable Corpora for Translation Studies

Iustina Ilisei, Diana Inkpen, Gloria Corpas, Ruslan Mitkov

Building comparable corpora for the investigation of translational hypotheses is an important task within the translation studies domain. This paper describes the compilation of a translational comparable corpus for the Romanian language. The resource comprises translated and non-translated news articles and it is designed to be used in the investigation of translational language and translational hypotheses.

Evaluation of a Bilingual Dictionary Extracted from Wikipedia

Angelina Ivanova

Machine-readable dictionaries play important role in the research area of computational linguistics. They gained popularity in such fields as machine translation and cross-language information extraction. Wiki-dictionaries differ dramatically from the traditional dictionaries: the recall of the basic terminology on the Mueller’s dictionary was 7.42%. Machine translation experiments with the Wiki-dictionary incorporated into the training set resulted in the rather small, but statistically significant drop of the the quality of the translation compared to the experiment without the Wiki-dictionary. We supposed that the main reason was domain difference between the dictionary and the corpus and got some evidence that on the test set collected from Wikipedia articles the model with incorporated dictionary performed better.

A Visualizing Annotation Tool for Semi-Automatical Building a Bilingual Corpus

Quoc Hung-Ngo, Werner Winiwarter

Bilingual corpora are critical resources for machine translation research and development since parallel corpora contain translation equivalences of various granularities. Manual annotation of word alignments is of significance to provide a gold-standard for developing and evaluating both example-based machine translation models and statistical machine translation models. The annotation process costs a lot of time and effort, especially with a corpus of millions of words. This paper presents research on using visualization for an annotation tool to build an English-Vietnamese parallel corpus, which is constructed for a Vietnamese-English machine translation system. We describe the specification of collecting data for the corpus, linguistic tagging, bilingual annotation, and the tools specifically developed for the manual annotation. An English-Vietnamese bilingual corpus of over 800,000 sentence pairs and 10,000,000 English words as well as Vietnamese words has been collected and aligned at the sentence level; and a part of this corpus containing 200 news articles was aligned manually at the word level.

SMT systems for less-resourced languages based on domain-specific data

Lene Offersgaard, Dorte Haltrup Hansen

In this paper we show that good SMT systems for less-resourced languages can be obtained by using even small amounts of high quality domain-specific data. We suggest a method to filter newly collected data for parallel corpora, using the internal alignment scores from the aligning process. The filtering process is easy to use and is based on open-source tools. The domain-specific data are used in combination with other public available resources for training SMT systems. Automatic evaluation shows that relatively small amounts of newly collected domain-specific data result in systems with promising BLEU scores in the range of 52.9 to 60.9. The LetsMT! platform is used to create the presented machine translation systems, where the flexible platform allows uploading the user's own data for training. The paper shows that the platform is a promising way of making SMT systems available for less-resourced languages.

Towards a Wikipedia-extracted Alpine Corpus

Magdalena Plamada, Martin Volk

This paper describes a method for extracting parallel sentences from comparable texts. We present the main challenges in creating a German-French corpus for the Alpine domain. We demonstrate that it is difficult to use the Wikipedia categorization for the extraction of domain-specific articles from Wikipedia, therefore we introduce an alternative information retrieval approach. Sentence alignment algorithms were used to identify semantically equivalent sentences across the Wikipedia articles. Using this approach, we create a corpus of sentence-aligned Alpine texts, which is evaluated both manually and automatically. Results show that even a small collection of extracted texts (approximately 10 000 sentence pairs) can partially improve the performance of a state-of-the-art statistical machine translation system. Thus, the approach is worth pursuing on a larger scale, as well as for other language pairs and domains.

Using Comparable Corpora to Track Diachronic and Synchronic Changes in Lexical Density and Lexical Richness

Sanja Štajner, Ruslan Mitkov

This study from the area of language variation and change is based on exploitation of the comparable diachronic and synchronic corpora of 20th century British and American English language (the ‘Brown family’ of corpora). We investigate recent changes of lexical density and lexical richness in two consecutive thirty-year time gaps in British English (1931–1961 and 1961–1991) and in 1961–1992 in American English. Furthermore, we compare the diachronic changes between these two language varieties and discuss the results of the synchronic comparison of these two features between British and American parts of the corpora (in 1961 and in 1991/2). Additionally, we explore the possibilities of these comparable corpora by using two different approaches to their exploitation: using the fifteen fine-grained text genres, and using only the four main text categories. Finally, we discuss the impact of the chosen approaches in making hypotheses about the way language changes.

Mining for Term Translations in Comparable Corpora

Dan Ştefănescu

This paper presents the techniques currently developed at RACAI for extracting parallel terminology from the comparable collection of Romanian and English documents collected in the ACCURAT project. Apart from being used for enriching translation models, parallel terminology can be (and very often is) a goal in itself, since such resources can be used for building dictionaries or indexing technical or domain-restricted documents.

Accurate phrase alignment in a bilingual corpus for EBMT systems

George Tambouratzis, Michalis Troullinos, Sokratis Sofianopoulos, Marina Vassiliou

An ongoing trend in the creation of Machine Translation (MT) systems concerns the automatic extraction of information from large bilingual parallel corpora. As these corpora are expensive to create, the largest possible amount of information needs to be extracted in a consistent manner. The present article introduces a phrase alignment methodology for transferring structural information between languages using only a limited-size parallel corpus. This is used as a first processing stage to support a phrase-based MT system that can be readily ported to new language pairs. The essential language resources used in this MT system include a large monolingual corpus and a small parallel one. An analysis of different alignment cases is provided and the solutions chosen are described. In addition, the application of the system to different language pairs is reported and the results obtained are compared across language pairs to investigate the language-independent aspect of the proposed approach.

Using Czech-English Parallel Corpora in Automatic Identification of *It*

Kateřina Veselovská, Nguy Giang Linh, Michal Novák

In this paper we have two goals. First, we want to present a part of the annotation scheme of the recently released Prague Czech-English Dependency Treebank 2.0 related to the annotation of personal pronoun *it* on the tectogrammatical layer of sentence representation. Second, we introduce experiments with the automatic identification of English personal pronoun *it* and its Czech counterpart. We design sets of tree-oriented rules and on the English side we combine them with the state-of-the-art statistical system that altogether results in an improvement of the identification. Furthermore, we design and successfully apply rules, which exploit information from the other language.

CLIR- and ontology-based approach for bilingual extraction of comparable documents

Manuela Yapomo, Gloria Corpas, Ruslan Mitkov

The exploitation of comparable corpora has proven to be a valuable alternative to rare parallel corpora in various Natural Language Processing tasks. Therefore many researchers have stressed the need for large quantities of such corpora and the scarcity of works on their compilation. This paper describes a CLIR-based method for automatic extraction of French-English comparable documents. At the start of the process, source documents are translated and most representative terms are extracted. The resulting keyword list is further enlarged with synonyms on the assumption that keyword expansion might improve the retrieval of such documents. Retrieval is performed on the indexed target collection and a further filtering step based mainly on temporal information and document length takes place. Preliminary results suggest that the employment of ontology could improve the performance of the system.

Poster Session

Saturday 26 May 2012, 15:35 – 16:30
Chairperson: Marko Tadić

Oral Presentations 3: Lexicon Extraction and Corpus Analysis

Saturday 26 May 2012, 16:30 – 17:50
Chairperson: Andrejs Vasiļjevs

ICA for Bilingual Lexicon Extraction from Comparable Corpora

Amir Hazem, Emmanuel Morin

Independent component analysis (ICA) is a statistical method used to discover hidden features from a set of measurements or observed data so that the sources are maximally independent. This paper reports the first results on using ICA for the task of bilingual lexicon extraction from comparable corpora. We introduce two representations of data using ICA. The first one is called global ICA (GICA) used to design a global representation of a context according to all the target entries of the bilingual lexicon, the second one is called local ICA (LICA) and is used to capture local information according to target bilingual lexicon entries that only appear in the context vector of the candidate to translate. Then, we merge both GICA and LICA to obtain our final model (GLICA). The experiments are conducted on two different corpora. The French-English specialised corpus ’breast cancer’ of 1 million words and the French-English general corpus ’Le Monde / New-York Times’ of 10 million words. We show that the empirical results obtained with GLICA are competitive with the standard approach traditionally dedicated to this task.

Improving Compositional Translation with Comparable Corpora

Hiroyuki Kaji, Takashi Tsunakawa, Yoshihoro Komatsubara

We improved the compositional term translation method by using comparable corpora. A bilingual lexicon consisting of pairs of word sequences within terms and their correlations is derived from a bilingual document-aligned corpus. Then, for an input term, compositional translations are produced together with their confidence scores by consulting the corpus-derived bilingual lexicon. Thus, we can select the correct translation for the input term from among as many candidate ones as possible. An experiment with a comparable corpus of Japanese and English scientific-paper abstracts demonstrated that compositional translation using the corpus-derived bilingual lexicon outperforms that using an ordinary bilingual lexicon. Future work includes the incremental improvement of the bilingual lexicon with correlations, the refinement of the confidence score, and the extension of the compositional translation model to allow word order to be changed.

Multi-word term extraction from comparable corpora by combining contextual and constituent clues

Nikola Ljubešić, Špela Vintar, Darja Fišer

In this paper we present an approach to automatically extract and align multi-word terms from an English-Slovene comparable health corpus. First, the terms are extracted from the corpus for each language separately using a list of user-adjustable morphosyntactic patterns and a term weighting measure. Then, the extracted terms are aligned in a bag-of-equivalents fashion with a seed bilingual lexicon. In the extension of the approach we also show that the small general seed lexicon can be enriched with domain-specific vocabulary by harvesting it directly from the comparable corpus, which significantly improves the results of multi-word term mapping. While most previous efforts in bilingual lexicon extraction from comparable corpora have focused on mapping of single words, the proposed technique successfully augments them in that it is able to deal with multi-word terms as well. Since the proposed approach requires minimal knowledge resources, it is easily adaptable for a new language pair or domain, which is one of its biggest advantages.

Textual Characteristics of Different-sized Corpora

Robert Remus, Mathias Bank

Recently, textual characteristics, i.e. certain language statistics, have been proposed to compare corpora originating from different genres and domains, to give guidance in language engineering processes and to estimate the transferability of natural language processing algorithms from one corpus to another. However, until now it is unclear how these textual characteristics behave for different-sized corpora. We monitor the behavior of 7 textual characteristics across 4 genres – news articles, Wikipedia articles, general web text and fora posts – and 10 corpus sizes, ranging from 100 to 3,000,000 sentences. Thereby we show, certain textual characteristics are almost constant across corpus sizes and thus might be used to reliably compare different-sized corpora, while others are highly corpus size-dependent and thus may only be used to compare similar- or same-sized corpora. Moreover we find, although textual characteristics vary from genre to genre, their behavior for increasing corpus size is quite similar.

EEOP2012: Exploring and Exploiting Official Publications

Sunday May 27, 2012

ABSTRACTS

Editor:

Steven Krauwer

Workshop Programme

Sunday May 27, 2012

09:00-09:05 Welcome and introduction by Steven Krauwer

09:05-09:50 **Invited talk:** Maarten Marx
Open Official Documents: Requirements and Opportunities

09:50-10:10 Michael Rosner and Andrew Attard
Intelligent Exploitation of Local Government Resources

10:10-10:30 Maria Palmerini, Ruben Cerolini, Giulio Santini and Francesco Cutugno
From Recording to Retrieving: A Proposal of a Complete System for Semi-automatic Reporting for Local and National Governments

10:30-11:00 Coffee break

11:00-11:20 Vidas Daudaravicius
Automatic Multilingual Annotation of EU Legislation with Eurovoc Descriptors

11:20-11:40 Francesca Frontini, Carlo Aliprandi, Clara Bacciu, Roberto Bartolini, Andrea Marchetti, Enrico Parenti, Fulvio Piccinonno and Tiziana Soru
GLOSS, an Infrastructure for the Semantic Annotation and Mining of Documents in the Public Security Domain

11:40-12:00 Oliver Mason, Aleksander Trklja and Dominik Vajn
Requirement Extraction from Transport Policy Documents

12:00-12:50 Open discussion
What are possible actions that could be undertaken to enhance the exploration and exploitation of official publications at the international, cross-national and national level?

12:50-13:00 Winding up and closing by Steven Krauwer

Workshop Organizers and Programme Committee

Steven Krauwer	Utrecht University / CLARIN ERIC
Ralf Steinberger	European Commission – Joint Research Centre (JRC)
Arjan van Hessen	University of Twente / CLARIN-NL
Nicoletta Calzolari	CNR-ILC / ELRA
Hans Uszkoreit	DFKI / META-NET

Introduction

The EEOP2012 workshop is dedicated to the exploitation and exploration of official publications in digital format, both at the international level (often multilingual) and at the national level (mostly monolingual, but in some cases multilingual as well). These publications can be in written, spoken or visual form or combinations thereof (e.g. written proceedings of parliaments, legislative documents, audio or video recordings of parliament sessions, simultaneous translations by interpreters or in sign language).

The workshop covers the whole lifecycle of these publications, ranging from acquisition, annotation, instrumentation, exploration of data and content, exploitation of data and content to support research and the development of tools and applications.
The main objectives of the workshop are:

- To create awareness of the importance of official publications by showing the research and development possibilities they offer;
- To share results, experiences and problems emerging from work on a variety of corpora, modalities and languages;
- To identify actions that could be undertaken to enhance the exploration and exploitation of official publications at the international, cross-national and national level.

Official publications can be of tremendous importance for the research communities interested in human language technology (in the broadest possible sense) and for the communities interested in linguistics, psychology, history, social sciences and political sciences because they have a number of specific characteristics that make them different from other language resources:

- If they exist in digital form they are normally public and free;
- They grow continuously;
- They are often multilingual and parallel;
- They lend themselves for exploitation (as training material for tools and sometimes possibly even for niche applications);
- They lend themselves for exploration to support linguistic studies, studies about human behaviour, about changes in society, attitudes, and many other possible research topics in the humanities and social sciences;
- Because of their comparability they lend themselves for porting technologies, methods and expertise between languages;
- They lend themselves for educational purposes for technologists, linguists and other scholars.

Primary audience of this workshop is:

- Language and speech technology researchers from academia and industry;
- Humanities and social sciences scholars with an interest in digital methods;
- Educators in these fields.

Additional beneficiaries, not necessarily present at LREC 2012:

- Professionals interested in analysing political behaviour or processes (e.g. journalists, policy makers, policy watchers);
- Parties interested in providing or exploiting such analysis tools on a commercial basis;
- Translation studies scholars;
- Comparative linguists.

Acknowledgement: This workshop was supported by the European Commission – Joint Research Centre (JRC)

Exploring and Exploiting Official Publications

Sunday May 27, 09:00-13:00
Oral session

Open Official Documents: Requirements and Opportunities *(Invited talk)*

Maarten Marx

In this talk we show results on a survey of the quality of the Parliamentary data in a number of European countries. We measure quality using the (Linked) Open Government Data requirements set out by the W3C eGov working group. We also show opportunities for information science researchers given by Open Government Data, focusing on Parliamentary data.

Intelligent Exploitation of Local Government Resources

Michael Rosner and Andrew Attard

Malta is divided into sixty-eight local councils each contributing to the most basic form of local government. Several meetings take place during which the councillors gather to discuss the maintenance and embellishment of the locality, each of which are noted down in Maltese. This paper concerns a corpus of local government documents. We suggest an approach to the problem of developing an intelligent browsing system that offers improved access to the information, for example to assist local councils in decision making, or to give members of the public more transparent way to browse local council documentation.

From Recording to Retrieving: A Proposal of a Complete System for Semi-automatic Reporting for Local and National Governments

Maria Palmerini, Ruben Cerolini, Giulio Santini and Francesco Cutugno

The system we present here gives the possibility of bringing multimediality into the process of information retrieval from audio, video and Italian texts derived by parliament reports. The aim is not only to improve and increase the different ways the official documents can be watched and listened to and retrieved, but also to let all this information be available for different categories of users. Cedat 85 has produced a web service thought to satisfy the requests of the Basilicata Region Council and Verona Town Council, but that, given its premises, aims to be applicable in a wider range of parliamentary environments.

Automatic Multilingual Annotation of EU Legislation with Eurovoc Descriptors

Vidas Daudaravicius

Automatic document annotation from a controlled conceptual thesaurus is useful for establishing precise links between similar documents. This study presents a language independent document annotation system based on features derived from a collocation segmentation method. Using the multilingual conceptual thesaurus EuroVoc, we evaluate the method, comparing it against other language independent methods based on single words and bigrams. Testing the method against the manually tagged multilingual corpus Acquis Communautaire 3.0 (AC) using all descriptors found there, we attain improvements in keyword assignment precision from 50.7 to 57.6 percent over three diverse languages (English, Lithuanian and Finnish) tested. We found high correlation between automatic assignment precision against document length and language features such as inflectiveness and compounding.

GLOSS, an Infrastructure for the Semantic Annotation and Mining of Documents in the Public Security Domain

Francesca Frontini, Carlo Aliprandi, Clara Bacciu, Roberto Bartolini, Andrea Marchetti, Enrico Parenti, Fulvio Piccinonno and Tiziana Soru

Efficient access to information is crucial in the work of organizations that require decision taking in emergency situations. This paper gives an outline of GLOSS, an integrated system for the analysis and retrieval of data in the environmental and public security domain. We shall briefly present the GLOSS infrastructure and its use, and how semantic information of various kinds is integrated, annotated and made available to the final users.

Requirement Extraction from Transport Policy Documents

Oliver Mason, Aleksander Trklja and Dominik Vajn

Requirements are an important concept in systems engineering. We present an approach to the automatic extraction of requirement statements from transport policy documents using a local grammar. The grammar has been developed using standard corpus linguistic methods. With a fairly straight forward grammar we can identify instances of requirements, as they are expressed using a small number of distinct surface patterns. One additional complication involved in this research is the issue of clearly separating requirements from policies or strategies, which are generally at a higher level. We have identified a number of different patterns that can help in distinguishing between such statements

4th International Workshop on Corpora for Research on EMOTION SENTIMENT & SOCIAL SIGNALS ES³ 2012

26 May 2012

ABSTRACTS

ES³ 2012
4th International Workshop on
Corpora for Research on EMOTION SENTIMENT & SOCIAL SIGNALS
Satellite of LREC 2012, ELRA, Full Day Workshop on Saturday, 26 May 2012, Istanbul, Turkey, http://emotion-research.net/sigs/speech-sig/es12

Editors:

Laurence Devillers, Björn Schuller, Anton Batliner, Paolo Rosso, Ellen Douglas-Cowie, Roddy Cowie, Catherine Pelachaud

Workshop Programme

09:00 – 09:10
Laurence Devillers
Opening

ORAL SESSION 1: MULTILINGUAL SENTIMENT RESOURCES AND ANALYSIS
(Chair: Paolo Rosso)

09:10 – 09:30
Andrea Gianti, Cristina Bosco, Viviana Patti, Andrea Bolioli and Luigi Di Caro
Annotating Irony in a Novel Italian Corpus for Sentiment Analysis

09:30 – 09:50
Sandi Pohorec, Ines Ceh, Peter Kokol and Milan Zorman
Sentiment Analysis Resources for Slovene Language

09:50 – 10:10
Ozan Cakmak, Abe Kazemzadeh, Dogan Can, Serdar Yildirim and Shrikanth Narayanan
Root-Word Analysis of Turkish Emotional Language

10:10 – 10:30
Minlie Huang, Lei Fang and Xiaoyan Zhu
A Chinese Corpus for Sentiment Analysis

10:30 - 11:00
COFFEE BREAK

ORAL SESSION 2: LAUGHTER AND SOCIAL SIGNAL PROCESSING
(Chair: Catherine Pelachaud)

11:00 – 11:20
Khiet Truong and Jürgen Trouvain
Laughter Annotations in Conversational Speech Corpora – Possibilities and Limitations for Phonetic Analysis

11:20 – 11:40
Radoslaw Niewiadomski, Jérôme Urbain, Catherine Pelachaud and Thierry Dutoit
Finding out the Audio and Visual Features that Influence the Perception of Laughter Intensity and Differ in Inhalation and Exhalation Phases

11:40 – 12:00
Gary McKeown, Roddy Cowie, Will Curran, Willibald Ruch and Ellen Douglas-Cowie
ILHAIRE Laughter Database

12:00 – 12:20
Jürgen Trouvain and Khiet Truong
Comparing Non-Verbal Vocalisations in Conversational Speech Corpora

12:20 – 12:40
Magalie Ochs, Paul Brunet, Gary McKeown, Catherine Pelachaud and Roddy Cowie
Smiling Virtual Characters Corpora

12:40 – 13:00
Isabella Poggi, Francesca D'Errico and Laura Vincze
Ridiculization in Public Debates: Making Fun of the Other as a Discrediting Move

13:00 – 14:00
LUNCH BREAK

ORAL SESSION 3: EMOTION AND AFFECT
(Chair: Laurence Devillers)

14:00 – 14:20
Rene Altrov and Hille Pajupuu
Estonian Emotional Speech Corpus: Theoretical Base and Implementation

14:20 – 14:40
Dipankar Das, Soujanya Poria, Chandra Mohan Dasari and Sivaji Bandyopadhyay
Building Resources for Multilingual Affect Analysis – A Case Study on Hindi, Bengali and Telugu

14:40 – 15:00
Clément Chastagnol and Laurence Devillers
Collecting Spontaneous Emotional Data for a Social Assistive Robot

15:00 – 15:20
Wenjing Han, Haifeng Li, Lin Ma, Xiaopeng Zhang and Björn Schuller
A Ranking-based Emotion Annotation Scheme and Real-life Speech Database

15:20 – 15:40
John Snel, Alexey Tarasov, Charlie Cullen and Sarah Jane Delany
A Crowdsourcing Approach to Labelling a Mood Induced Speech Corpora

15:40 – 16:00
Serkan Özkul, Elif Bozkurt, Shahriar Asta, Yücel Yemez and Engin Erzin
Multimodal Analysis of Upper-Body Gestures, Facial Expressions and Speech

16:00 – 16:30
COFFEE BREAK

ORAL SESSION 4: CROSS-DISCIPLINE PERSPECTIVES
(Chair: Haifeng Li)

16:30 – 16:50
Dietmar Rösner, Jörg Frommer, Rico Andrich, Rafael Friesen, Matthias Haase, Manuela Kunze, Julia Lange and Mirko Otto
LAST MINUTE: a Novel Corpus to Support Emotion, Sentiment and Social Signal Processing

16:50 – 17:10
Katia Lida Kermanidis
Mining Authors' Personality Traits from Modern Greek Spontaneous Text

17:10 – 17:30
Antonio Reyes and Paolo Rosso
Building Corpora for Figurative Language Processing: The Case of Irony Detection

17:30 – 17:50
Marcela Charfuelan and Marc Schröder
Correlation Analysis of Sentiment Analysis Scores and Acoustic Features in Audiobook Narratives

17:50 – 18:10
Effie Mouka, Voula Giouli, Aggeliki Fotopoulou and Ioannis E. Saridakis
Opinion and Emotion in Movies: a Modular Perspective to Annotation

18:10 – 18:30
Closing Discussion
Paolo Rosso

19:00 – 21:00
Optional Common Dinner
Covered by participants

Editors

Laurence Devillers	Université Paris-Sorbonne 4, France
Björn Schuller	Technische Universität München, Germany
Anton Batliner	Friedrich-Alexander-University, Germany
Paolo Rosso	Universitat Politècnica de Valencia, Spain
Ellen Douglas-Cowie	Queen's University Belfast, UK
Roddy Cowie	Queen's University Belfast, UK
Catherine Pelachaud	CNRS - LTCI, France

Workshop Organizers/Organizing Committee

Laurence Devillers	Université Paris-Sorbonne 4, France
Björn Schuller	Technische Universität München, Germany
Anton Batliner	Friedrich-Alexander-University, Germany
Paolo Rosso	Universitat Politècnica de Valencia, Spain
Ellen Douglas-Cowie	Queen's University Belfast, UK
Roddy Cowie	Queen's University Belfast, UK
Catherine Pelachaud	CNRS - LTCI, France

Workshop Programme Committee

Vered Aharonson	AFEKA, Israel
Alexandra Balahur	EC’s Joint Research Centre, Italy
Felix Burkhardt	Deutsche Telekom, Germany
Carlos Busso	University of Texas at Dallas, USA
Rafael Calvo	University of Sydney, Australia
Erik Cambria	National University Singapore, Singapore
Antonio Camurri	University of Genova, Italy
Mohamed Chetouani	Université Paris 6, France
Thierry Dutoit	University of Mons, Belgium
Julien Epps	University of New South Wales, Australia
Anna Esposito	IIASS, Italy
Hatice Gunes	Queen Mary University, UK
Catherine Havasi	MIT Media Lab, USA
Bing Liu	University of Illinois at Chicago, USA
Florian Metze	Carnegie Mellon University, USA
Shrikanth Narayanan	University of Southern California, USA
Maja Pantic	Imperial College London, UK
Antonio Reyes	Universidad Politècnica de Valencia, Spain
Fabien Ringeval	Université de Fribourg, Switzerland
Peter Robinson	University of Cambridge, UK
Florian Schiel	LMU, Germany
Jianhua Tao	Chinese Academy of Sciences, China
José A. Troyano	Universidad de Sevilla, Spain
Tony Veale	University College Dublin, Ireland
Alessandro Vinciarelli	University of Glasgow, UK
Haixun Wang	Microsoft Research Asia, China

Preface

The fourth instalment of the workshop series on Corpora for Research on Emotion held at LREC aims at further cross-fertilisation between the highly related communities of emotion and affect processing based on acoustics of the speech signal, and linguistic analysis of spoken and written text, i.e., the field of sentiment analysis including figurative languages such as irony, sarcasm, satire, metaphor, parody, etc. At the same time, the workshop opens up for the emerging field of behavioural and social signal processing including signals such as laughs, smiles, sighs, hesitations, consents, etc. Besides data from human-system interaction, dyadic and human-to-human data, its labelling and suited models as well as benchmark analysis and evaluation results on suited and relevant corpora were invited. By this, we aim at bridging between these larger and highly connected fields: Emotion and sentiment are part of social communication, and social signals are highly relevant in helping to better understand affective behaviour and its context. For example, understanding of a subject's personality is needed to make better sense of observed emotional patterns. At the same time, non-linguistic behaviour such as laughter and linguistic analysis can give further insight into the state or personality trait of the subject.

All these fields further share a unique trait: Genuine emotion, sentiment and social signals are hard to collect, ambiguous to annotate, and tricky to distribute due to privacy reasons. In addition, the few available corpora suffer from a number of issues owing to the peculiarity of these young and emerging fields: As in no related task, different forms of modelling exist, and ground truth is never solid due to the often highly different perception of the mostly very few annotators. Due to data sparseness, cross-validation without strict partitioning including development sets and without strict separation of speakers and subjects throughout partitioning are frequently seen.

Laurence Devillers, Björn Schuller, Anton Batliner, Paolo Rosso,
Ellen Douglas-Cowie, Roddy Cowie, Catherine Pelachaud

SESSION 1: MULTILINGUAL SENTIMENT RESOURCES AND ANALYSIS

Saturday 26 May, 9:10 – 10:30
Chairperson: Paolo Rosso

Paper Title

Annotating Irony in a Novel Italian Corpus for Sentiment Analysis

Andrea Gianti, Cristina Bosco, Viviana Patti, Andrea Bolioli and Luigi Di Caro

Abstract

In this paper we describe our current work on Senti–TUT, a novel Italian corpus for sentiment analysis. This resource includes annotations concerning both sentiment and morpho-syntax, in order to make available several possibilities of further exploitation related to sentiment analysis. For what concerns the annotation at sentiment level, we focus on irony and we selected therefore texts on politics from a social media, namely Twitter, where irony is usually applied by humans. Our aim is to add a new sentiment dimension, which explicitly accounts for irony, to a sentiment analysis classification framework based on polarity annotation. The paper describes the data set, the features of the annotation both at sentiment and morpho-syntactic level, the procedures and tools applied in the annotation process. Finally, it shows the preliminary experiments we are carrying on in order to validate the annotation work.

Sentiment Analysis Resources for Slovene Language

Sandi Pohorec, Ines Ceh, Peter Kokol and Milan Zorman

Abstract

Slovene language lacks resources for sentiment analysis of natural language. Several large lexical resources are available, but they only provide information on word lemmas and part-of-speech tags. This paper presents an experiment in which the well-known General Inquirer (GI) dictionary has been automatically translated into Slovene with the use of several multilingual resources. We have implemented an automated system for the translation of the General Inquirer dictionary as well as processed large amounts of Slovene text in order to provide the basic statistical data, used for language recognition, in the form of n-gram distributions. Each word entry in the translated dictionary has been lemmatized and each entry provides all known word forms. The resources presented here offer the capability to automatically detect if the text is in Slovene language and analyze the content with GI regardless of the word form.

Root-Word Analysis of Turkish Emotional Language

Ozan Cakmak, Abe Kazemzadeh, Dogan Can, Serdar Yildirim and Shrikanth Narayanan

Abstract

This paper describes a model for the perceived emotion of Turkish sentences based on the emotions associated with the constituent words. In our model, each emotion is mapped to a point in the continuous space defined by three emotional attributes: valence, activation, and dominance. We

collected a large data set through two independent surveys: a word-level survey that prompted users with emotional words and asked them to assign each word a continuous emotional interval, and a sentence-level survey that prompted users with emotional sentences collected from 31 children's books and asked them to rate each sentence on a discrete emotional scale. The word-level survey was aimed at creating a core affective lexicon for Turkish. It is difficult to build a comprehensive affective lexicon for Turkish due to its very productive morphology that generates a very large vocabulary. We deal with the sparsity issues caused by the large word vocabulary by analyzing the emotional content of word roots. Our experimental results indicate that there is a strong correlation between the emotions attributed to Turkish word roots and the Turkish sentences.

A Chinese Corpus for Sentiment Analysis

Minlie Huang, Lei Fang and Xiaoyan Zhu

Abstract

Sentiment analysis and opinion mining has been a hot topic in the text mining and natural language processing communities. There have been a number of corpora in English or other western languages, either for sentiment classification, or for opinion extraction. However, to the best of our knowledge, few Chinese counterparts exist for these opinion mining tasks. In this paper, we introduce a Chinese corpus for opinion mining. The corpus contains two parts: a set of multi-domain sentences, with sentiment polarity annotated, and a set of multi-domain aspect-opinion pairs and corresponding polarities, which were obtained automatically from almost 5 million custom reviews. We present the corpus statistics, annotation guidelines, and discussions of how to use the corpus. We believe that such a corpus is potentially useful for sentence-level sentiment classification, aspect-level opinion extractions, opinion summarization, and so on.

SESSION 2: LAUGHTER AND SOCIAL SIGNAL PROCESSING

Saturday 26 May, 11:00 – 13:00
Chairperson: Catherine Pelachaud

Laughter Annotations in Conversational Speech Corpora – Possibilities and Limitations for Phonetic Analysis

Khiet Truong and Jürgen Trouvain

Abstract

Existing laughter annotations provided with several publicly available conversational speech corpora (both multiparty and dyadic conversations) were investigated and compared. We discuss the possibilities and limitations of these rather coarse and shallow laughter annotations. There are definition issues to be considered with respect to speech-laughs and the segmentation of laughs: what constitutes one laugh, and when does a laugh start and end? Despite these issues, some durational and voicing analyses can be performed. We found for all corpora considered that overlapping laughs are longer in duration and are generally more voiced than non-overlapping laughs. For a finer-grained acoustic analysis, we find that a manual re-labeling of the laughs adhering to a more standardized laughter annotations protocol would be optimal.

Finding out the Audio and Visual Features that Influence the Perception of Laughter Intensity and Differ in Inhalation and Exhalation Phases

Radoslaw Niewiadomski, Jérôme Urbain, Catherine Pelachaud and Thierry Dutoit

Abstract

This paper presents the results of the analysis of laughter expressive behavior. First we present the intensity annotation study of an audiovisual corpus of spontaneous laughter. In the second part of the paper we present the analysis of audio and visual cues that influence the perception of laughter intensity, as well as the study of audio and visual features that differ in laughter inhalation and exhalation phases.

ILHAIRE Laughter Database

Gary McKeown, Roddy Cowie, Will Curran, Willibald Ruch and Ellen Douglas-Cowie

Abstract

The ILHAIRE project seeks to scientifically analyse laughter in sufficient detail to allow the modelling of human laughter and subsequent generation and synthesis of laughter in avatars suitable for human machine interaction. As part of the process an incremental database is required providing different types of data to aid in modelling and synthesis. Here we present an initial part of that database in which laughs were extracted from a number of pre-existing databases. Emphasis as been placed on extraction of laughs that are social and conversational in style as there are already existing databases that include instances of hilarious laughter. However, an attempt has been made to exhaustively extract all instances of laughter from databases that were not designed for the purpose of generating hilarious laughter. Theses databases are: the Belfast Naturalistic Database, the HUMAINE Database, the Green Persuasive Database, the Belfast Induced Natural Emotion Database and the SEMAINE Database.

Comparing Non-Verbal Vocalisations in Conversational Speech Corpora

Jürgen Trouvain and Khiet Truong

Abstract

Conversations do not only consist of spoken words but they also consist of non-verbal vocalisations. Since there is no standard to define and to classify (possible) non-speech sounds the annotations for these vocalisations differ very much for various corpora of conversational speech. There seems to be agreement in the six inspected corpora that hesitation sounds and feedback vocalisations are considered as words (without a standard orthography). The most frequent non-verbal vocalisation are laughter on the one hand and, if considered a vocal sound, breathing noises on the other.

Smiling Virtual Characters Corpora

Magalie Ochs, Paul Brunet, Gary McKeown, Catherine Pelachaud and Roddy Cowie

Abstract

To create smiling virtual characters, the different morphological and dynamic characteristics of the virtual characters smiles and the impact of the virtual characters smiling behavior on the users need to be identified. For this purpose, we have collected two corpora: one directly created by users and the other resulting from the interaction between virtual characters and users. We present in details these two corpora in the article.

Ridiculization in Public Debates: Making Fun of the Other as a Discrediting Move

Isabella Poggi, Francesca D'Errico and Laura Vincze

Abstract

The paper analyzes acts of ridiculization in public debates. Ridiculization is a communicative act that, through expressing a negative evaluation of lack of power on some person, makes her feel abased, isolated, dropped out of the group, thus fulfilling a function of moralistic aggression and one of enhancing group identity. Ridiculization is seen here as a way to discredit the opponent in a political debate, in such a way as to make him/her less credible and less persuasive in front of the audience. Several cases of ridiculization are presented, and the cues to ridiculization acts are listed, from smile and laughter to simple serious words, to pretended compassion and praise, to irony, imitation and parody.

SESSION 3: EMOTION AND AFFECT

Saturday 26 May, 14:00 – 16:00
Chairperson: Paolo Rosso

Estonian Emotional Speech Corpus: Theoretical Base and Implementation

Rene Altrov and Hille Pajupuu

Abstract

The establishment of the Estonian Emotional Speech Corpus (EESC) began in 2006 within the framework of the National Programme for Estonian Language Technology at the Institute of the Estonian Language. The corpus contains 1,234 Estonian sentences that express anger, joy and sadness, or are neutral. The sentences come from text passages read out by non-professionals who were not given any explicit indication of the target emotion. It was assumed that the content of the text would elicit an emotion in the reader and that this would be expressed in their voice. This avoids the exaggerations of acted speech. The emotion of each sentence in the corpus was then determined by listening tests. The corpus is publicly available at http://peeter.eki.ee:5000/. This article gives an overview of the theoretical starting-points of the corpus and their usefulness for its implementation.

Building Resources for Multilingual Affect Analysis – A Case Study on Hindi, Bengali and Telugu

Dipankar Das, Soujanya Poria, Chandra Mohan Dasari and Sivaji Bandyopadhyay

Abstract

The rapid growth of affective texts in the Web 2.0 and multilingualism in search engines motivate us to prepare the emotion/affect data for three Indian languages (Hindi, Bengali and Telugu). This paper reports the development of the WordNet Affects and SemEval 2007 affect sensing corpora in three target Indian languages from the available English sources that were provided in the Affective Text shared task on the SemEval 2007 workshop. The linguistic evaluation on the developed resources proposed various morals from the perspective of affect analysis in the target languages. Two emotion analysis systems, baseline systems followed by morphology driven systems have been developed and the evaluation results of the systems produce satisfactory results in comparison with the English and Japanese. The rapid growth of affective texts in the Web 2.0 and multilingualism in search engines motivate us to prepare the emotion/affect data for three Indian languages (Hindi, Bengali and Telugu). This paper reports the development of the WordNet Affects and SemEval 2007 affect sensing corpora in three target Indian languages from the available English sources that were provided in the Affective Text shared task on the SemEval 2007 workshop. The linguistic evaluation on the developed resources proposed various morals from the perspective of affect analysis in the target languages. Two emotion analysis systems, baseline systems followed by morphology driven systems have been developed and the evaluation results of the systems produce satisfactory results in comparison with the English and Japanese.

Collecting Spontaneous Emotional Data for a Social Assistive Robot

Clément Chastagnol and Laurence Devillers

Abstract

The French ARMEN ANR-funded project aims at building an assistive robot for elderly and disabled people. This robot is controlled by a VCA (Virtual Conversational Agent), interacting with the subjects in a natural, spoken fashion. We focus in this paper on the gathering of emotional data in interaction with the VCA (or her voice only in the first gathering). 77 patients have participated in the data collection. The data will be used for building an emotion detection system. The specific difficulty in this project lies in the large variety of user voices (elderly, pathological) and affective behaviors of the patient. A questionnaire on the acceptability of the VCA and the quality of the interaction is also analysed and shows that the interaction with the VCA was deemed positive by the subjects.

A Ranking-based Emotion Annotation Scheme and Real-life Speech Database

Wenjing Han, Haifeng Li, Lin Ma, Xiaopeng Zhang and Björn Schuller

Abstract

In this paper, we propose employing a learning-to-rank algorithm to the recognition of emotion in speech, and construct a novel ranking-based speech emotion recognition (SER) framework. We firstly design a ranking-based annotation scheme to collect high-reliability labels for model training. Next, we use the ranking scores to measure speakers' emotions and apply a learning-to-rank algorithm called ListNet to recognise (i.e., rank) emotion. A linear neural network is then trained for SER. Furthermore, a reference-based emotion visualisation approach is proposed to describe speakers' emotion fluctuation relative to a normal situation. Finally, feasibility of these methods is validated on the medium-scaled Mandarin real-life emotion corpus introduced for the first time which features massive 300 k individual pair wise comparisons.

A Crowdsourcing Approach to Labelling a Mood Induced Speech Corpora

John Snel, Alexey Tarasov, Charlie Cullen and Sarah Jane Delany

Abstract

This paper demonstrates the use of crowdsourcing to accumulate ratings from naive listeners as a means to provide labels for a naturalistic emotional speech dataset. In order to do so, listening tasks are performed with a rating tool, which is delivered via the web. The rating requirements are based on the classical dimensions, activation and evaluation, presented to the participant as two discretised 5-point scales. Great emphasis is placed on the participant's overall understanding of the task, and on the ease-of-use of the tool so that labelling accuracy is reinforced. The accumulation process is ongoing with a goal to supply the research community with a publicly available speech corpus.

Multimodal Analysis of Upper-Body Gestures, Facial Expressions and Speech

Serkan Özkul, Elif Bozkurt, Shahriar Asta, Yücel Yemez and Engin Erzin

Abstract

We propose a multimodal framework for correlation analysis of upper body gestures, facial expressions and speech prosody patterns of a speaker in spontaneous and natural conversation. Spontaneous upper body, face and speech gestures exhibit a broad range of structural relationships and have not been previously analyzed together to the best of our knowledge. In this study we present a multimodal database of spontaneous conversation. Within this database to identify cross modal correlations, we first perform unsupervised temporal segmentation of each modality using hidden Markov model (HMM) structures. Then, we perform correlation analysis based on mutual information measure, which is experimentally computed over the joint histograms of recurrent temporal segments (patterns) of modalities.

SESSION 4: CROSS-DISCIPLINE PERSPECTIVES

Saturday 26 May, 16:30 – 18:10
Chairperson: Paolo Rosso

LAST MINUTE: a Novel Corpus to Support Emotion, Sentiment and Social Signal Processing

Dietmar Rösner, Jörg Frommer, Rico Andrich, Rafael Friesen, Matthias Haase, Manuela Kunze, Julia Lange and Mirko Otto

Abstract

We present the LAST MINUTE corpus, a multimodal data collection taken from a carefully designed WoZ experiment that allows to investigate how users interact with a companion system in a mundane situation with the need for planning, re-planning and strategy change. The experiments have been performed with a cohort of N = 130 subjects, balanced in age, gender and educational level. The resulting corpus does not only comprise high quality recordings from audio, video and psychobiological signals, it contains as well transcripts from all interactions and data from a battery of well established psychometric questionnaires filled out by all subjects. For a subgroup of subjects audio records and transcripts from an additional post hoc interview are available as well.

Mining Authors’ Personality Traits from Modern Greek Spontaneous Text

Katia Lida Kermanidis

Abstract

The present work describes the automatic recognition of authors’ personality traits, based on the linguistic properties of their writing. An SVM classifier is applied for the first time to Modern Greek textual features in order to learn the Big Five personality traits for the author. Linguistic features are limited to low-level morphological categorical features. Cross-language findings, even if still premature, are interesting, as several correlations between linguistic properties and personality traits that hold for English, seem to apply in Modern Greek as well. Bootstrapping helps towards improving classification accuracy and avoiding the problem of overfitting.

Building Corpora for Figurative Language Processing: The Case of Irony Detection

Antonio Reyes and Paolo Rosso

Abstract

Figurative language is one of the most arduous topics that natural language processing (NLP) has to face. Unlike literal language, the former takes advantage of linguistic devices, such as metaphor, analogy, ambiguity, irony, sarcasm, and so on, in order to communicate more complex meanings, which usually represent a serious problem, not only for computers, but for humans as well. In this article we describe the problem of figurative language processing concerning corpus-based approaches. This type of language is quite common in web contents; however, its automatic processing entails a huge challenge, both theoretically as pragmatically. Here we describe the need of automatically building training corpora with objective and reliable data. In this respect, we are focused on addressing a quite complex device: irony. Such linguistic phenomenon, which is widespread in web content, has important implications for tasks such as sentiment analysis, opinion mining, or even advertising.

Correlation Analysis of Sentiment Analysis Scores and Acoustic Features in Audiobook Narratives

Marcela Charfuelan and Marc Schröder

Abstract

We investigate possible correlations between sentiment analysis scores obtained for sentences of Mark Twain’s novel “The Adventures of Tom Sawyer” and acoustic features extracted from the same sentences in the corresponding audiobook. We have found that scores derived from movie reviews or categorisation of emotional stories seem to be more close to the acoustics in the narrative, in particular more correlated with average energy and mean fundamental frequency (F0). We have designed an experiment intended to predict the levels of acoustic expressivity in arbitrary text using sentiment analysis scores and the number of words in the text.

Opinion and Emotion in Movies: a Modular Perspective to Annotation

Effie Mouka, Voula Giouli, Aggeliki Fotopoulou and Ioannis E. Saridakis

Abstract

This paper presents an ongoing effort work focusing on the development of an audiovisual corpus resource and its annotation in terms of sentiments and opinions. A modular annotation schema has been employed based on the specifications of existing schemas and extending or adapting them to cater for the peculiarities of the corpus-specific data.

Joint ISA-7, SRSL-3 and I2MRT Workshop on Interoperable Semantic Annotation

26 May 2012

ABSTRACTS

Editors:

Harry Bunt, Manuel Alcantara-Plá, Peter Wittenburg

Workshop Programme

08.30 – 08:30 Registration
08:45 - 09:00 Workshop Opening

09:00 - 10:30 **Session: Semantic representation and multimodal resources**
09:00 - 09:30 Mehdi Manshadi and James Allen: *A Universal Representation for Shallow and Deep Semantics*
09:30 - 10:00 Rodolfo Delmonte and Agata Rotondi: *Treebanks of Logical Forms: They are Useful Only if Consistent*
10:00 - 10:30 Hennie Brugman and Mark Lindeman: *A Publication Platform for Open Annotations*

10:30 - 11:00 coffee break

11:00 - 13:00 **Session: Annotation of spatial information**
11:00 - 11:30 James Pustejovsky, Jessica Moszkowics and Marc Verhagen: *The Current Status of ISO-Space*
11:30 - 12:00 Robert Gaizauskas, Emma Barker, Ching-Lan Chang, Leon Derczynski, Michael Phiri and Chengzhi Peng: *Applying ISO-Space to Healthcare Facility Design Evaluation Reports*
12:00 - 12:30 Antje Müller: *Location and Path - Annotating Sense of the German Prepositions "auf" and "über"*
12:30 - 13:00 Linda Meini, Giovanna Marotta, Leonardo Lenci and Margherita Donati: *An XML Annotation Scheme for Space in an Italian Corpus*

13:00 - 14:00 lunch break

14:00 - 16:00 **Session: Semantic Roles and their annotation**
14:00 - 14:30 *Project ISO-Semantic Roles* (Martha Palmer)
14:30 - 15:00 Claire Bonial, Weston Feely, Jena Hwang and Martha Palmer: *Empirically Validating VerbNet using SemLink*
15:00 - 16:00 *The Lexlink project* (Collin Baker, Christiane Fellbaum, Martha Palmer)

16:00 - 16:30 tea break

16:30 - 18:00 **Session: Interoperable semantic annotation in ISO projects**
16:30 - 17:00 Kiyong Lee: *Interoperable Spatial and Temporal Annotation Schemes*
17:00 - 17:30 Harry Bunt, Rashmi Prasad and Aravind Joshi: *First Steps Towards an ISO Standard for Annotating Discourse Relations*
17:30 - 18:00 *Project ISO-Basics: Principles of Semantic Annotation* (Harry Bunt)
18:00 Workshop Closing

Workshop Organizers

Harry Bunt	Tilburg University
Manuel Alcantara-Plá	Universidad Autónoma de Madrid
Peter Wittenburg	Max Planck Institute for Psycholinguistics, Nijmegen
Thierry Declerck	DFKI, Saarbrücken
Dafydd Gibbon	University of Bielefeld
Nancy Ide	Vassar College, Poughkeepsie, NY
Steven Krauwer	Universiteit Utrecht
Kiyong Lee	Korea University, Seoul
Lorenza Mondada	Université de Lyon 2
James Pustejovsky	Brandeis University, Waltham, MA
Laurent Romary	INRIA/Humboldt Universität Berlin
Oliver Schreer	Fraunhofer Institute for Telecomuunications, Berlin

Workshop Programme Committee

Jan Alexandersson	DFKI, Saarbrücken
Stefan Baumann	Universität Köln
Jonas Beskow	KTH, Stockholm
Paul Buitelaar	National University of Ireland, Galway
Harry Bunt	Tilburg University
Thierry Declerck	DFKI, Saarbrücken
Raquel Fernandez Rovira	Universiteit van Amsterdam
Anette Frank	Universität Heidelberg
Dafydd Giboon	Universität Bielefeld
Koiti Hasida	AIST, Tokyo
Nancy Ide	Vassar College, Poughkeepsie, NY
Michael Kipp	University of Applied Sciences, Augsburg
Kiyong Lee	Korea University, Seoul
Inderjeet Mani	Chiang Mai, Thailand
Jean-Clause Martin	LIMSI, Orsay
Lorenza Mondada	Université de Lyon 2
Martha Palmer	University of Colorado, Boulder
Volha Petukhova	Vicomtech, San Sebastian
Andrei Popescu-Belis	Idiap, Martigny, Switzerland
Rarhmi Prasad	University of Wisconsin, Milwaukee
James Pustejovsky	Brandeis University, Wlatham, MA
Laurent Romary	INRIA/Humboldt Universität Berlin
Oliver Schreer	Fraunhofer Institute for Telecomuunications, Berlin
Mark Steedman	University of Edinburgh
Mariët Theune	Universiteit Twente
Isabel Trancoso	INESC, Lisbon

Introduction

Three initiatives have joined forces in this workshop, which is concerned with issues in semantic annotation for language resources, especially in relation to spoken and multimodal language data, and with the interoperability and integration of resources and tools.

ISA-7 is the Seventh Workshop on Interoperable Semantic Annotation, and forms part of a series of workshops of ISO TC 37/SC 4 (Language Resources) jointly with ACL-SIGSEM (Computational Semantics). These workshops bring together experts in the annotation of semantic information as expressed in text, speech, gestures, graphics, video, images, and in multiple modalities combined. Examples of semantic annotation include the markup of events, time, space, dialogue acts, discourse relations, and semantic roles, for which the ISO organization pursues the establishment of annotation standards, in order to support the creation of interoperable semantic resources.

SRSL-3 is the Third Workshop on Semantic Representation of Spoken Language in Speech and Multimodal Corpora. In these workshops researchers convene who are working on speech and multimodal resources for the semantic annotation of related corpora, and take their inspiration from the observation that the semantic gap between the content conveyed by speech and other modalities and their formal representation is a burning issue in a range of tasks such as content mining, information extraction, dialogue processing, interactive story-telling, assisted health care,and human-robot interaction.

I2MRT (Integration and Interoperability for Multimodal Resources and Tools) is an initiative to address infrastructure aspects of the creation and use of interoperable multimodal resources. Main objectives of I2MRT are to create awareness of the need to make multimodal data visible via standardized methods and accessible via registered data centers; to discuss possibilities of harmonization and standardization of multimodal annotation schemes and possible mappings between schemes; to discuss ways to make cutting-edge technologies available to multimodality researchers; and to build a community that is committed to work further on these issues.

Harry Bunt
Manuel Alcantar-Plá
Peter Wittenburg

Session Semantic Representation and Multimodal Resources

Saturday 26 May, 9:00 – 10:30
Chairperson:

A Universal Representation for Shallow and Deep Semantics

Mehdi Manshadi and James Allen

We define a graphical semantic representation that readily captures the partial semantic analyses produced by shallow processing techniques, yet is also as fully expressive as the representations used in deep analysis systems, including discourse processing. While in most existing natural language systems, robustness often comes at the expense of shallowness, our representation is designed to bridge this gap. The framework is not specific to a particular semantic theory, and may be translated into various target languages. In particular, the translation into first order or intentional logic is transparent. We show how the framework is able to capture more complex semantic phenomena, such as scopal adverbials and predicate modifiers. The graphical frameworks allows us to define mathematical notions to determine the well-formedness of a representation or the coherence of the corresponding sentence once we have the complete semantic representation of a sentence. A unique property of our semantic framework is to encode some syntactic properties of a sentence as well. We define an evaluation framework for this formalism that allows one to compute semantic recall and precision measures given gold standard representations. e

Treebanks of Logical Forms: They are Useful Only if Consistent

Rodolfo Delmonte and Agata Rotondi

Logical Forms are an exceptionally important linguistic representation for highly demanding semantically related tasks like Question/ Answering and Text Understanding, but their automatic production at runtime is higly error-prone. The use of a tool like XWNet and other similar resources would be beneficial for all the NLP community, but not only. The problem is: Logical Forms are useful as long as they are consistent, otherwise they would be useless if not harmful. Like any other resource that aims at providing a meaning representation, LFs require a big effort in manual checking order to reduce the number of errors to the minimum acceptable – less than 1% - from any digital resource. As will be shown in detail in the paper, the available resources – XWNet, WN30-lfs, ILF - suffer from lack of a careful manual checking phase, and the number of errors is too high to make the resource usable as is. We classified mistakes by their syntactic or semantic type in order to facilitate a revision of the resource that we intend to do using regular expressions. We also commented extensively on semantic issues and on the best way to represent them in Logical Forms.

A Publication Platform for Open Annotations

Hennie Brugman and Mark Lindeman

Abstract The OpenAnnotation Consortium introduced a generic model for representing annotations of resources and resource segments that complies to principles of the World Wide Web and Linked Data. This paper introduces a platform for storing, retrieving, searching, exchanging, harvesting and publishing Open Annotations on the web. It describes design considerations, functionality and architecture. Our Open Annotation server platform is set up as a distributed system with server instances that can exchange annotations in a peer-to-peer way. Each instance can persistently

publish annotations using principles of the web and thereby adds 'annotatability' to annotations themselves and to annotation 'Bodies'. Additionally, the annotation platform provides efficient search and implements a Dashboard for server management tasks.
The web-oriented nature of the platform raises a number of interesting issues and opportunities that are discussed in some depth. For example, in general uploaded annotations do not have resolvable http URIs. Assigning those is not trivial. Indexing strategy, determining the boundaries of an annotation in an RDF graph and searching for annotations whose Body is somewhere else on the web are other issues that are discussed.

Session Annotation of spatial information

Saturday 26 May, 11:00 – 13:00
Chairperson:

The Current Status of ISO-Space

James Pustejovsky, Jessica L. Moszkowicz and Marc Verhagen

We report on ISO-Space version 1.4, an annotation specification for capturing spatial and spatiotemporal information in natural language that is now in its fourth incarnation. This version substantially improves upon earlier ISO-Space specifications in a few notable ways. The representation of locations is no longer overloaded such that geolocations have a more complete annotation and non-geolocations are captured with specific tags. In addition, interactions with existing annotation standards such as TimeML have been clarified. The treatment of spatial prepositions has been modified so that their annotation is more suggestive of what spatial relationships should hold between two spatial objects. Finally, spatial relationships are now captured with four distinct link tags: qualitative spatial links for topological relationships, orientation links for non-topological relations, movement links for motion, and measure links for detailing a metric relationship between two spatial objects or what the dimensions of a particular object are. The most recent version of the specification is presented with illustrative examples. We conclude with some outstanding issues that have yet to be captured in the specification.

Applying ISO-Space to Healthcare Facility Design Reports

Robert Gaizauskas, Emma Barker, Ching-Lan Chang, Leon Derczynski, Michael Phiri and Chengzhi Peng

This paper describes preliminary work on the spatial annotation of textual reports about healthcare facility design to support the long-term goal of linking report content to a three-dimensional building model.
Emerging semantic annotation standards enable formal description of multiple types of discourse information. In this instance, we investigate the application of a spatial semantic annotation standard at the building-interior level, where most prior applications have been at inter-city or street level.
Working with a small corpus of design evaluation documents, we have begun to apply the ISO-Space specification to annotate spatial information in healthcare facility design evaluation reports.
These reports present an opportunity to explore semantic annotation of spatial language in a novel situation.
We describe our application scenario, report on the sorts of spatial language found in design evaluation reports, discuss issues arising when ISO-Space is applied to building-level entities, and propose possible extensions to ISO-Space to address the issues encountered.

Location and Path – Annotating Senses of the German Prepositions "auf" and "über"

Antje Müller

Many difficulties concerning so-called spatial prepositions arise from an insufficient subclassification of the prepositionsí interpretations. Since there is no one-to-one mapping from possible locations to prepositions there is a substantial need to differentiate the diverse interpretations of one preposition. In this paper we present an approach for a subclassification of some spatial prepositions. We will focus on the correlation between the German route prepositions *über* and *durch* and their static local counterparts *auf* and *in*.

Route prepositions are often considered to be decomposable in a PATH function and a location. We will show that this assumption plus an adequate description of the underlying location results in a systematic classification of preposition senses. It is useful for the annotation of spatial preposition senses as well as for the analyses of the interpretations. For annotation the spatial interpretations are organized in a categorization tree. On the way through the tree different features are picked up that determine the respective interpretation. So every interpretation can be characterized as a set of features paired with the form of the preposition. This set-theoretic view of interpretations makes semantic relations between different interpretations of one and the same prepositions as well as between related interpretations of different prepositions apparent.

An XML Annotation Scheme for Space in an Italian Corpus

Linda Meini, Giovanna Marotta, Leonardo Lenci, and Margherita Donati

The new resource we present consists of a corpus of oral spatial descriptions performed by congenital blind and sighted Italian subjects. The collection of the data is part of a wider project on semantic representations in the language of the blind, carried out at the Department of Linguistics, University of Pisa. The long-term goal of the project is to use the evidence collected on congenital blind subjects to get at a better understanding of the relationship between linguistic and perceptive information. The corpus is currently being enhanced with different layers of annotation, focusing on spatial information. The annotation allows us to highlight the effect of the specific lexical and grammatical features of Italian on the encoding of space (e.g. with respect to the way spatial relations are encoded in motion verbs). Our resource is not only one of the few annotated corpora of spoken Italian, but it is also the first one that focuses on spatial categories.

Session: Semantic Role Annotation and Definition

Saturday 26 May, 14:00 – 16:00
Chairperson:

Empirically Validating VerbNet Using SemLink

Claire Bonial, Weston Feely, Jena D. Hwang and Martha Palmer

This research describes efforts to empirically validate a lexical resource, VerbNet, using the PropBank annotations found in the SemLink corpus. As a test case, we examine the frequency with which verbs in SemLink appear in the Caused-Motion syntactic frame: NP-V-NP-PP (e.g., textit{She poured water into the bowl). To do this, we find the frequency with which a given verb is used in this construction, we then determine each verb's VerbNet class membership, and compare the overall frequency of the Caused-Motion construction in the verb class to how the verbs' behavior is currently represented in VerbNet. We find evidence that VerbNet's current classification fails to capture generalizations about the likelihood of a class' compatibility with the Caused-Motion construction. Specifically, classes where Caused-Motion is currently represented in VerbNet as a characteristic syntactic frame were found to have a lower frequency of realization in that frame than other classes where Caused-Motion is not represented. We therefore suggest augmenting VerbNet's classification with information on the probability that a class will participate in a certain syntactic frame, and given the challenges of this research, offer potential improvements for increasing the interoperability of VerbNet.

Session: ISO projects on semantic annotation

Date / Time *[Saturday 26 May, 16:30 – 18:00]*
Chairperson:

Interoperable Spatial and Temporal Annotation Schemes

Kiyong Lee

ISO-TimeML (2012) was just published as an international standard for the annotation of temporal and event-related information in language. Almost at the same time, Pustejovsky and Moszkowicz (2012) produced a revised version of ISO-Space specifications as a spatial annotation scheme. The purpose of this paper is to argue for the need of making these two annotation schemes interoperable to allow a unified treatment of annotating spatial and temporal information in language. This task is mainly motivated by many occurrences of spatio-temporal signals (e.g., at, in, after) in text that trigger both spatial and temporal relations between various types of basic elements annotated to text offsets or segments, called markables. We argue that these two semantic annotation schemes can be made interoperable by merging some of their specifications, especially concerning the use of spatial or temporal signals and those relations triggered by these signals and, furthermore, that this merging results in designing an integrated spatio-temporal annotation and interpretation scheme

First Steps Towards an ISO Standard for Annotating Discourse Relations

Harry Bunt, Rashmi Prasad and Aravind Joshi

This paper describes initial studies in the context of a new effort within ISO to design an international standard for the annotation of discourse with relations that account for its coherence, in particular so-called `discourse relations'. This effort takes the Penn Discourse Treebank (PDTB) as its starting point, and applies a methodology for defining semantic annotation languages which distinguishes an abstract syntax, defining annotation structures as set-theoretical constructs, a concrete syntax, that defines a reference XML-based format for representing annotation structures, and a formal semantics. A first attempt is described to formulate an abstract syntax and a concrete syntax for the annotation scheme underlying the PDTB. The abstract syntax clearly shows an overall structure for a general-purpose standard for annotating discourse relations, while the resulting concrete syntax is much more readable and semantically transparent than the original format. Moreover, some additional elements are introduced which have an optional status, making the proposed representation format compatible not only with the PDTB but also with other approaches.

The Third Workshop on Computational Models of Narrative

(CMN'12)

26-27 May 2012

ABSTRACTS

Editor:

Mark A. Finlayson

Workshop Programme

Saturday, 26 May 2012

12:30 Welcome and Introduction, *M.A. Finlayson*

13:00 **Invited Keynote**: Crowd Sourcing Narrative Logic: Towards a Computational Narratology with CLÉA, *J.C. Meister*

Session I: Representation

14:00 Toward Sequencing "Narrative DNA": Tale Types, Motif Strings and Memetic Pathways, *S. Darányi, P. Wittek, L. Forró*

14:20 Computational Models of Narratives as Structured Associations of Formalized Elementary Events, *G.P. Zarri*

14:35 Objectivity and Reproducibility of Proppian Narrative Annotations, *R. Bod, B. Fisseni, A. Kurji, B. Löwe*

14:50 An Experiment to Determine whether Clustering will Reveal Mythemes, *R. Lang, J.G. Mersch*

15:00 In Search of an Appropriate Abstraction Level for Motif Annotations, *F. Karsdorp, P. van Kranenburg, T. Meder, D. Trieschnigg, A. van den Bosch*

15:15 Understanding Objects in Online Museum Collections by Means of Narratives, *C. van den Akker, M. van Erp, L. Aroyo, R. Segers, L. van der Meij, S. Lêgene and G. Schreiber*

15:30 **Best Student Paper on a Cognitive Science Topic**: Indexter: A Computational Model of the Event-Indexing Situation Model for Characterizing Narratives, *R.E. Cardona-Rivera, B.A. Cassell, S.G. Ware, R.M. Young*

15:50 Coffee Break

16:30 Towards Finding the Fundamental Unit of Narrative: A Proposal for the Narreme, *A. Baikadi, R.E. Cardona-Rivera*

16:40 People, Places and Emotions: Visually Representing Historical Context in Oral Testimonies, *A.T. Chen, A. Yoon, R. Shaw*

Session II: Corpora

16:55 TrollFinder: Geo-Semantic Exploration of a Very Large Corpus of Danish Folklore, *P.M. Broadwell, T.R. Tangherlini*

17:15 A Hybrid Model and Memory Based Story Classifier, *B. Ceran, R. Karad, S. Corman, H. Davulcu*

17:30 A Crowd-Sourced Collection of Narratives for Studying Conflict, *R. Swanson and A. Jhala*

17:50 Towards a Culturally-Rich Shared Narrative Corpus: Suggestions for the Inclusion of Culturally Diverse Narrative Genres, *V. Romero, J. Niehaus, P. Weyhrauch, J. Pfautz, S.N. Rielly*

18:00 Towards a Digital Resource for African Folktales, *D.O. Ninan and O.A. Odejobi*

18:15 Formal Models of Western Films for Interactive Narrative Technologies, *B. Magerko, B. O'Neill*

18:35 End

Workshop Programme

(cont. . .)

Sunday, 26 May 2012

Session III: Similarity

9:00 Detecting Story Analogies from Annotations of Time, Action and Agency, *D.K. Elson*

9:20 Story Comparison via Simultaneous Matching and Alignment, *M.P. Fay*

9:35 Similarity of Narratives, *L. Michael*

9:55 Which Dimensions of Narratives are Relevant for Human Judgments of Story Equivalence?, *B. Fisseni, B. Löwe*

10:10 Story Retrieval and Comparison using Concept Patterns, *C.E. Krakauer, P.H. Winston*

10:30 Coffee Break

Session IV: Generation

11:00 From the Fleece of Fact to Narrative Yarns: A Computational Model of Composition, *P. Gervás*

11:20 Is this a DAG that I see before me?: An Onomasiological Approach to Narrative Analysis and Generation, *M. Levison, G. Lessard*

11:40 Automatically Learning to Tell Stories about Social Situations from the Crowd, *B. Li, S. Lee-Urban, D.S. Appling, M.O. Riedl*

12:00 Prototyping the Use of Plot Curves to Guide Story Generation, *C. León, P. Gervás*

12:10 Simulating Plot: Towards a Generative Model of Narrative Structure, *G.A. Sack*

12:30 Lunch

Session V: Persuasion

14:30 A Choice-Based Model of Character Personality in Narrative, *J.C. Bahamon, R.M. Young*

14:45 Persuasive Precedents, *F. Bex, T. Bench-Capon, B. Verheij*

15:00 Integrating Argumentation, Narrative and Probability in Legal Evidence, *B. Verheij*

15:10 Arguments as Narratives, *A. Wyner*

15:20 Towards a Computational Model of Narrative Persuasion: A Broad Perspective, *J. Niehaus, V. Romera, J. Pfautz, S.N. Reilly, R. Gerrig, P. Weyhrauch*

15:30 Discussion

16:00 Coffee Break

16:30 Discussion

18:00 End

Organizing Committee

Mark A. Finlayson (Chair)	Massachusetts Institute of Technology, USA
Pablo Gervás	Universidad Complutense de Madrid, Spain
Deniz Yuret	Koç University, Turkey
Floris Bex	University of Dundee, UK

Programme Committee

Steve Corman	Arizona State University, USA
Barbara Dancygier	University of British Columbia, Canada
Hasan Davulcu	Arizona State University, USA
David K. Elson	Google, USA
Matthew P. Fay	Massachusetts Institute of Technology, USA
Andrew Gordon	Institute for Creative Technologies, USA
Benedikt Löwe	University of Amsterdam, the Netherlands
Livia Polanyi	LDM Associates, USA
Emmett Tomai	University of Texas-Pan American, USA
Bart Verheij	University of Groningen, the Netherlands
Patrick H. Winston	Massachusetts Institute of Technology, USA
R. Michael Young	North Carolina State University, USA

Invited Keynote

Saturday 26 May, 13:00–14:00

13:00–14:00

Crowd Sourcing Narrative Logic: Towards a Computational Narratology with CLÉA

Jan Christoph Meister

I will discuss a collaborative, computer aided approach towards building and exploiting a shared resource that can aid further research into the history and development of narrative, as well as into its phenomenology and logic. The particular example illustrating this approach is a project called CLÉA, short for *Collaborative Literature Éxploration and Annotation*.

Session I: Representation

Saturday 26 May, 14:00–15:50 & 16:30–16:55

14:00–14:20

Toward Sequencing "Narrative DNA": Tale Types, Motif Strings and Memetic Pathways

Sándor Darányi, Peter Wittek, László Forró

The Aarne-Thompson-Uther Tale Type Catalog (ATU) is a bibliographic tool which uses metadata from tale content, called motifs, to define tale types as canonical motif sequences. The motifs themselves are listed in another bibliographic tool, the Aarne-Thompson Motif Index (AaTh). Tale types in ATU are defined in an abstracted fashion and can be processed like a corpus. We analyzed 219 types with 1202 motifs from the "Tales of magic" (types 300-749) segment to exemplify that motif sequences show signs of recombination in the storytelling process. Compared to chromosome mutations in genetics, we offer examples for insertion/deletion, duplication and, possibly, transposition, whereas the sample was not sufficient to find inverted motif strings as well. These initial findings encourage efforts to sequence motif strings like DNA in genetics, attempting to find for instance the longest common motif subsequences in tales. Expressing the network of motif connections by graphs suggests that tale plots as consolidated pathways of content help one memorize culturally engraved messages. We anticipate a connection between such networks and Waddington's epigenetic landscape.

14:20–14:35

Computational Models of Narratives as Structured Associations of Formalized Elementary Events

Gian Piero Zarri

In this paper, we describe the conceptual tools that, in an NKRL context (NKRL = Narrative Knowledge Representation Language), allow us to obtain a (computer-suitable) description of full "narratives" as logically- and temporally-ordered streams of formalized "elementary events." After having introduced, first, the main principles underpinning NKRL, we describe in some detail the characteristics of the second order (reification-based) tools, like the "completive construction" and the "binding occurrences," which implement concretely the association of the NKRL-formalized elementary events. Examples concerning some recent applications of NKRL in different domains will be used in the paper to better explain the use of these tools.

14:35–14:50

Objectivity and Reproducibility of Proppian Narrative Annotations

Rens Bod, Bernhard Fisseni, Aadil Kurji, Benedikt Löwe

A formal narrative representation is a procedure assigning a formal description to a natural language narrative. One of the goals of the *computational models of narrative* community is to understand this procedure better in order to automatize it. A formal framework fit for automatization should allow for objective and reproducible representations. In this paper, we present empirical work focussing on objectivity and reproducibility of the formal framework by Vladimir Propp (1928). The experiments consider Propp's formalization of Russian fairy tales and formalizations done by test subjects in the same formal framework; the data show that some features of Propp's system such as the assignment of the characters to the *dramatis personae* and some of the functions are not easy to reproduce.

14:50–15:00

An Experiment to Determine Whether Clustering Will Reveal Mythemes

R. Raymond Lang, John G. Mersch

Claude Levi-Strauss proposes a universal structure for narrative myths. The structure is expressed as the canonical formula, $f_x(a) : f_y(b) \approx f_x(b) : f_{a-1}(y)$, where the four terms of the formula denote bundles of gross constituent units, his term for predicate relations. The bundles are referred to as mythemes. The deep meaning of a myth is given by associating its semantic content with the terms of the formula. The analytic approach to myths is hindered by (1) circularity between the bundles and their components, and (2) heavy reliance on expert knowledge. This project is to develop a system for the algorithmic identification of these bundles. The investigation is starting with clustering of only word senses (semantemes) and will proceed to clustering of predicate relations. The number of desired clusters is known, and the clustered objects are non-numeric, so an appropriate algorithm is k-mediods, using distance metrics computed with the WordNet::Similarity Perl module. Status of the experiment and planned directions for the work are described.

15:00–15:15

In Search of an Appropriate Abstraction Level for Motif Annotations

Folgert Karsdorp, Peter van Kranenburg, Theo Meder, Dolf Trieschnigg, Antal van den Bosch

We present ongoing research on the role of motifs in oral transmission of stories. We assume that motifs constitute the primary building blocks of stories. On the basis of a quantitative analysis we show that the level of motif annotation utilized in the Aarne-Thompson-Uther folktale type catalogue is well suited to analyze two genres of folktales in terms of motif sequences. However, for the other five genres in the catalogue the annotation level is not apt, because it is unable to bring to front the commonalities between stories.

15:15–15:30

Understanding Objects in Online Museum Collections by Means of Narratives

Chiel van den Akker, Marieke van Erp, Lora Aroyo, Roxane Segers, Lourens van der Meij, Susan Lêgene, and Guus Schreiber

In this contribution, we present the narrative model used in Agora, an interdisciplinary project of the history and computer science departments at VU University Amsterdam and two cultural heritage institutions, the Rijksmuseum in Amsterdam and Sound & Vision in Hilversum. In the Agora project, we develop methods and techniques to support the narrative understanding of objects in online museum collections. A first demonstrator is now being tested. Here, our focus is on the specificity of modeling narratives in the heritage and history domain and the solutions Agora offers to specific problems of that domain.

In Agora, we believe that the interpretation of objects in online museum collections is supported by enriching the museum collection metadata with a structured notion of historical events and the (semi-)automatically generation of proto-narratives from those events. Starting from historical theory, three proto-narratives are distinguished: a biographical, a conceptual, and a topological proto-narrative. These proto-narratives are organizations of events based on the theory of narrative and historical theory. Proto-narratives not only take basic characteristics of the narrative into account, but also historical periods and complex historical events.

15:30–15:50

Best Student Paper on a Cognitive Science Topic
Indexter: A Computational Model of the Event-Indexing Situation Model for Characterizing Narratives

Rogelio E. Cardona-Rivera, Bradley A. Cassell, Stephen G. Ware, R. Michael Young

Previous approaches to computational models of narrative have successfully considered the internal coherence of the narratives structure. However, narratives are also externally focused and authors often design their stories to affect users in specific ways. In order to better characterize the audience in the process of modeling narrative, we introduce Indexter: a computational model of the Event-Indexing Situation Model, a cognitive framework which predicts the salience of previously experienced events in memory based on the current event the audience is experiencing. We approach computational models of narrative from a foundational perspective, and feel that salience is at the core of comprehension. If a particular narrative phenomenon can be expressed in terms of salience in a persons memory, the phenomenon, in principle, is representable in our model. This paper provides the fundamental bases of our approach as a springboard for future work which will use this model to reason about the audiences mental state, and to generate narrative fabula and discourse intended to achieve a specific narrative effect.

16:30–16:40

Towards Finding the Fundamental Unit of Narrative: A Proposal for the Narreme

Alok Baikadi, Rogelio E. Cardona-Rivera

Verb- and action-based event representations have been the cornerstone of narrative representation. However, these suffer from a lack of specificity as to the level of abstraction being discussed. For example, a single verb-based event can be elaborated *ad infinitum*, generating arbitrarily many new verb-based events. In this position paper, we present a proposal for the fundamental unit of narrative, which we call the *narreme*. Our contribution is two-fold. First, we present the structure of the narreme, which encodes the state of the narrative, not the state of the world. Second, we present the ways narremes can be combined, which gives rise to the structure of the narrative itself. These combinations have special properties which account for the causal, temporal and intentional relationships between the events that make up a narrative. Lastly, we present an interpretation of common narrative tasks within the context of the narreme.

16:40–16:55

People, Places and Emotions: Visually Representing Historical Context in Oral Testimonies

Annie T. Chen, Ayoung Yoon, Ryan Shaw

This paper presents visualizations to facilitate users ability to understand personal narratives in the historical and sociolinguistic context that they occurred. The visualizations focus on several elements of narrative time, space, and emotion to explore oral testimonies of Korean “comfort women,” women who were forced into sexual slavery by Japanese military during World War II. The visualizations were designed to enable viewers to easily spot similarities and differences in life paths among individuals and also form an integrated view of spatial, temporal and emotional aspects of narrative. By exploring the narratives through the interactive interfaces, these visualizations facilitate users’ understandings of the unique identities and experiences of the comfort women, in addition to their collective and shared story. Visualizations of this kind could be integrated into a toolkit for humanities scholars to facilitate exploration and analysis of other historical narratives, and thus serve as windows to intimate aspects of the past.

Session II: Corpora

Saturday 26 May, 16:55–18:20

16:55–17:15

TrollFinder: Geo-Semantic Exploration of a Very Large Corpus of Danish Folklore

Peter M. Broadwell, Timothy R. Tangherlini

We propose an integrated environment for the geo-navigation of a very large folklore corpus (>30,000 stories). Researchers of traditional storytelling are largely limited to existing indices for the discovery of stories. These indices rarely include geo-indexing, despite a fundamental premise of folkloristics that stories are closely related to the physical environment. In our approach, we develop a representation of latent semantic connections between stories and project these into a map-based navigation and discovery environment. Our preliminary work is based on the pre-existing corpus indices and a shared-keyword index, coupled to an index of geo-referenced places mentioned in the stories. Combining these allows us to produce heat maps of the relationship between places and a first level approximation of the story topics. The heat maps reveal concentrations of topics in a specific place. A researcher can use these topic concentrations as a method for building and refining research questions. We also allow for spatial querying, an approach that allows a researcher to discover topics that are particularly related to a specific place. Our corpus representation can be extended to include multimodal network representations of the corpus and LDA topic models to allow for additional visualizations of latent corpus topics.

17:15–17:30

A Hybrid Model and Memory Based Story Classifier

Betul Ceran, Ravi Karad, Steven Corman, Hasan Davulcu

A story is defined as “an actor(s) taking action(s) that culminates in a resolution(s).” In this paper, we investigate the utility of standard keyword based features, statistical features based on shallow-parsing (such as density of POS tags and named entities), and a new set of semantic features to develop a story classifier. This classifier is trained to identify a paragraph as a “story,” if the paragraph contains mostly story(ies). Training data is a collection of expert-coded story and non-story paragraphs from RSS feeds from a list of extremist web sites. Our proposed semantic features are based on suitable aggregation and generalization of <Subject, Verb, Object> triplets that can be extracted using a parser. Experimental results show that a model of statistical features alongside memory-based semantic linguistic features achieves the best accuracy with a Support Vector Machine (SVM) based classifier.

17:30–17:50

A Crowd-Sourced Collection of Narratives for Studying Conflict

Reid Swanson, Arnav Jhala

In this paper we have described a corpus that provides many of the details required to understand the context and dynamics that factor into a persons emotional state, the perception of consequences and their likely selection of a conflict resolution strategy. We believe this information is useful for meta-reasoning about narrative through a deeper knowledge about peoples thought process. It also provides enough detail for attempting a data driven approach for modeling the mental process of conflict resolution in computational agents that respond in a way that people find believable. Finally, we have provided a methodology to extend the corpus, so that over time we may cover a broader spectrum of conflicts and target specific domains when they are needed for applications.

17:50–18:00

Towards a Culturally-Rich Shared Narrative Corpus: Suggestions for the Inclusion of Culturally Diverse Narrative Genres

Victoria Romero, James Niehaus, Peter Weyhrauch, Jonathan Pfautz, Scott Neal Reilly

This paper proposes that the inclusion of culturally diverse narrative genres should be an explicit goal when developing a shared narrative corpus. We argue that narrative genres from underrepresented cultures and from cultures relevant to specific applications of computational narrative research should be prioritized. We offer the example of Mexican narcocorridos as a narrative genre that satisfies both these criteria.

18:00–18:15

Towards a Digital Resource for African Folktales

Deborah O. Ninan, Odetunji A. Ọdẹ́jọbí

This paper explores the development of a digital resource that is amenable to the formal specification of African folktales. The ultimate aim of this project is to develop computational structure and models for the narrative underlying African Folktale. We collected a number of Yorùbá folktales with corresponding English translations. We then analysed their components and the structure of the narrative that they embodied. The requirements of a markup language to capture the content and structure of the narratives was proposed. Ongoing work is aimed at the development of a framework for the computational model and the automatic generation of folktales based on this data.

18:15–18:35

Formal Models of Western Films for Interactive Narrative Technologies

Brian Magerko and Brian O'Neill

Interactive narrative technologies have typically addressed the authoring bottleneck problem by focusing on authoring a tractable story space (i.e. the space of possible experiences for a user) coupled with an AI technology for mediating the users journey through this space. This article describes an alternative, potentially more general and expressive approach to interactive narrative that focuses on the procedural representation of story construction between an AI agent and a human interactor. This notion of procedural interaction relies on shared background knowledge between all actors involved; therefore, we have developed a body of background knowledge for improvising Western-style stories that includes the authoring of scripts (i.e. prototypical joint activities in Westerns). This article describes our methodology for the design and development of these scripts, the formal representation used for encoding them in our interactive narrative technology, and the lessons learned from this experience in regards to building a story corpus for interactive narrative research.

Session III: Similarity

Sunday 27 May, 9:00–10:30

9:00–9:20

Detecting Story Analogies from Annotations of Time, Action and Agency

David K. Elson

We describe the Story Intention Graph (SIG) as a model of narrative meaning that is amenable to both corpus annotation and computational inference. The relations, focusing on time, action and agency, can express a range of thematic scenarios and lend themselves to the automatic detection of story similarity and analogy. An evaluation finds that such detection outperforms a propositional similarity metric in predicting human judgments of story similarity in the Aesop domain.

9:20–9:35

Story Comparison via Simultaneous Matching and Alignment

Matthew P. Fay

Story understanding is an essential piece of human intelligence. If we are to develop artificial intelligence with the cognitive capacities of humans, our systems must not only be able to understand stories but also to incorporate them into the thought process as humans do. The techniques I present enable efficient gap filling through story alignment. The approach demonstrated leverages the solid foundation of bio-informatics alignment techniques to create the simultaneous matching and alignment algorithm for story comparison. The algorithm provides a large improvement in efficiency in solving the matching problem, reducing the search space from 1030 nodes to 535 nodes in an example narrative comparison. The technique enables effective story comparison as an important step towards enabling higher level narrative intelligence.

9:35–9:55

Similarity of Narratives

Loizos Michael

The task of recognizing narrative similarity is put forward as a concrete metric of success for machine narrative understanding. For this task, one seeks to determine which of two narratives is more similar to a third target narrative. As a first step towards building machines that achieve this goal, we investigate herein the notion of narrative similarity through a computational lens. We approach similarity as a balancing act between a listeners search for commonalities between stories, and an authors quest to guard a storys intended inferences.

9:55–10:10

Which Dimensions of Narratives are Relevant for Human Judgments of Story Equivalence?

Bernhard Fisseni, Benedikt Löwe

We present an experimental approach to determining natural dimensions of story comparison. The results show that untrained test subjects generally do not privilege structural information. When asked to justify sameness ratings, they may refer to content, but when asked to state differences, they mostly refer to style, concrete events, details and motifs. We conclude that adequate formal models of narratives must represent such non-structural data.

10:10–10:30

Story Retrieval and Comparison using Concept Patterns

Caryn E. Krakauer, Patrick H. Winston

Traditional story comparison uses key words to determine similarity. However, the use of key words misses much of what makes two stories alike. The method we have developed use high level concept patterns, which are comprised of multiple events, and compares them across stories. Comparison based on concept patterns can note that two stories are similar because both contain, for example, revenge and betrayal concept patterns, even though the words revenge and betrayal do not appear in either story, and one may be about kings and kingdoms while the other is about presidents and countries. Using a small corpus of 15 conflict stories, we have shown that similarity measurement using concept patterns does, in fact, differ substantially from similarity measurement using key words. The Goldilocks principle states that features should be of intermediate size; they should be not too big, and they should not too small. Our work can be viewed as adhering to the Goldilocks principle because concept patterns are features of intermediate size, hence not so large as an entire story, because no story will be exactly like another story, and not so small as individual words, because individual words tend to be common in all stories taken from the same domain. While our goal is to develop a human competence model, we note application potential in retrieval, prediction, explanation, and grouping.

Session IV: Generation
Sunday 27 May, 11:00–12:30

11:00–11:20
From the Fleece of Fact to Narrative Yarns: A Computational Model of Composition
Pablo Gervás

From a given observable set of events, a large number of stories may be composed, by deciding to select or omit specific events, by restricting attention to smaller subsets of the overall setting, by focusing on particular characters, or by narrating the chosen events in different order. This particular task of narrative composition is not covered by existing models of storytelling or cognitive accounts of the writing task. This paper presents a model of the task of narrative composition as a set of operations that need to be carried out to obtain a span of narrative text from a set of events that inspire the narration. To provide guidance in structuring the task, an analogy is drawn between the narrative composition task and that of manufacturing textile fibres, with corresponding concepts of heckling the original material into fibres, then twisting these fibres into richer and better yarns. The model explores a set of intermediate representations required to capture the structure that is progressively imposed on the material, and connects this content planning task with a classic pipeline for natural language generation. As an indicative case study, an initial implementation of the model is applied to a chess game understood as a formalised set of events susceptible of story-like interpretations. The relationships between this model and existing models from other fields (narratological studies, cognitive accounts of writing, AI models of story generation, and natural language generation architectures) is discussed.

11:20–11:40
"Is this a DAG that I see before me?" sAn Onomasiological Approach to Narrative Analysis and Generation
Michael Levison, Greg Lessard

We present a framework for the analysis of literary texts by means of a semantic representation based on the use of directed acyclic graphs which may be threaded in various ways to represent elements of plot, character perspective, narrative sequencing and setting. The model is illustrated by application to a simple fairy tale and to a Sherlock Holmes story. We argue that it is possible to represent in this way, in a manner accessible to non-computer scientists, the high-level dependencies which underlie a text as well as particular characteristics of literary texts, including the use of various recurring narrative sequences. We provide examples of the functional representation used, of the graphical representations achieved and the results obtained when the semantic representations are used to drive a natural language generator.

11:40–12:00
Automatically Learning to Tell Stories about Social Situations from the Crowd
Boyang Li, Stephen Lee-Urban, Darren Scott Appling, and Mark O. Riedl

Narrative intelligence is the use of narrative to make sense of the world and to communicate with other people. The generation of stories involving social and cultural situations (eating at a restaurant, going on a date, etc.) requires an extensive amount of experiential knowledge. While this knowledge can be encoded in the form of scripts, schemas, or frames, the manual authoring of these knowledge structures presents a significant bottleneck in the creation of systems demonstrating narrative intelligence. In this paper we describe a technique for automatically learning robust, script-like knowledge from crowdsourced narratives. Crowdsourcing, the use of anonymous human workers, provides an opportunity for rapidly acquiring a corpus of highly specialized narratives about sociocultural situations. We describe a three-stage approach to script acquisition and learning. First, we query human workers to write natural language narrative examples of a given situation. Second, we learn the set of possible events that can occur in a situation by finding semantic similarities between the narrative examples. Third, we learn the relevance of any event to the situation and extract a probable temporal ordering between events. We describe how these scripts, which we call plot graphs, can be utilized to generate believable stories about social situations.

12:00–12:10

Prototyping the Use of Plot Curves to Guide Story Generation

Carlos León, Pablo Gervás

Setting objectives for automatic story generation is needed for a story generation system to produce content. Among the potentially useful methods, curves defining the evolution of specific features of a narrative that evolve along time are particularly appropriate because they focus on the evolution of those features and are easy to create, modify and understand by human users. In this paper we propose a theoretical definition of curve-based story generation, its relation to existing story generation algorithms and how this theory can be applied to new systems.

12:10–12:30

Simulating Plot: Towards a Generative Model of Narrative Structure

Graham Alexander Sack

This paper explores the application of computer simulation techniques to the fields of literary studies and narratology by developing a model for plot structure and characterization. Using a corpus of 19th Century British novels as a case study, the author begins with a descriptive quantitative analysis of character names, developing a set of stylized facts about the way narratives allocate attention to their characters. The author shows that narrative attention in many novels appears to follow a long tail distribution. The author then constructs an explanatory model in NetLogo, demonstrating that basic assumptions about plot structure are sufficient to generate output consistent with the real novels in the corpus.

Session V: Persuasion

Sunday 27 May, 14:30–15:30

14:30–14:45

A Choice-Based Model of Character Personality in Narrative

Julio César Bahamón, R. Michael Young

The incorporation of interesting and compelling characters is one of the key components of effective narrative. Well-developed characters have features that enable them to significantly enhance the believability and overall quality of a story. In this paper we present preliminary research on the development of a computational model aimed at facilitating the inclusion of compelling characters in narrative that is automatically generated by a planning-based system. The model centers on the use of an intelligent process to express character personality. In this model, personality is operationalized as behavior that results from choices made by a character in the course of a story. This operationalization is based on the Big Five personality structure and results from behavioral psychology studies that link behavior to personality traits. We hypothesize that the relationship between choices and the actions they lead to can be used in narrative to produce the perception of specific personality traits in an audience.

14:45–15:00

Persuasive Precedents

Floris Bex, Trevor Bench-Capon, Bart Verheij

Stories can be a powerful vehicle of persuasion. We typically use stories to link known events into coherent wholes. One way to establish coherence is to appeal to past examples, real or fictitious. These examples can be chosen and critiqued using legal case-based reasoning (CBR) techniques. In this paper, we apply these techniques to factual stories, assessing a story about the facts using precedents. We thus show how legal reasoning in a CBR model is equally applicable to reasoning with factual stories.

15:00–15:10

Integrating Argumentation, Narrative and Probability in Legal Evidence

Bart Verheij

Reasoning on the basis of legal evidence has been analysed using three types of approaches: argumentative, narrative and probabilistic. As each type of approach has been defended as a complete account of evidential reasoning, it is natural to assume that there is an integrating perspective. It is here proposed that a logico-probabilistic argumentation theory can integrate argumentative, narrative and probabilistic approaches to legal evidence.

15:10–15:20

Arguments as Narratives

Adam Wyner

Aspects of narrative coherence are proposed as a means to investigate and identify arguments from text. Computational analysis of argumentation largely focuses on representations of arguments that are either abstract or are constructed from a logical (e.g. propositional or first order) knowledge base. Argumentation schemes have been advanced for stereotypical patterns of defeasible reasoning. While we have well-formedness conditions for arguments in a first order language, namely the patterns for inference, the conditions for argumentation schemes is an open question, and the identification of arguments in the wild is problematic. We do not understand the source of rules from which inference follows; formally, well-formed arguments can be expressed even with random sentences; moreover, argument indicators are sparse, so cannot be relied upon to identify arguments. As automated extraction of arguments from text increasingly finds important applications, it is pressing to isolate and integrate indicators of argument. To specify argument well-formedness conditions and identify arguments from unstructured text, we suggest using aspects of narrative coherence.

15:20–15:30

Towards a Computational Model of Narrative Persuasion: A Broad Perspective

James Niehaus, Victoria Romero, Jonathan Pfautz, Scott Neal Reilly, Richard Gerrig, Peter Weyhrauch

This paper presents a preliminary view on the elements of persuasive narratives from a computational perspective. We argue for a broad perspective of narrative persuasion, drawing on existing literature from multiple disciplines. We present a brief, first-steps analysis of the possible narrative elements that may influence narrative persuasion. Finally, we consider how these elements may influence the formation of narrative corpora.

@NLP can u tag #user_generated_content ?! via lrec-conf.org

26 May 2012

ABSTRACTS

Editor:

Maite Melero

Workshop Programme

14:00 – 14:15 Welcome and introduction

14:15 – 14:40 Óscar Muñoz-García and Carlos Navarro (Havas Media), *Comparing user generated content published in different social media sources*

14:40 – 15:05 Mehdi Aminian, Tetske Avontuur, Zeynep Azar, Iris Balemans, Laura Elshof, Rose Newell, Nanne van Noord, Alexandros Ntavelos, Menno van Zaanen (Tilburg University), *Assigning part-of-speech to Dutch tweets*

15:05 – 15:30 Diana Maynard, Kalina Bontcheva, Dominic Rout (University of Sheffield), *Challenges in developing opinion mining tools for social media*

15:30 – 16:00 Alejandro Mosquera and Paloma Moreda (University of Alicante), *A Qualitative Analysis of Informality Levels In Web 2.0 Texts: The Facebook Case Study*

16:00 – 16:30 Coffee break

16:30 – 16:55 Joan Codina and Jordi Atserias (Fundació Barcelona Media), *What is the text of a Tweet?*

16:55 – 17:20 Jennifer Garland, Stephanie Strassel, Safa Ismael, Zhiyi Song, Haejoong Lee (Linguistic Data Consortium, University of Pennsylvania), *Linguistic Resources for Genre-Independent Language Technologies: User-Generated Content in BOLT*

17:20 – 18:00 Panel Discussion

Workshop Organizers

Laura Alonso i Alemany	Universidad Nacional de Córdoba (Argentina)
Jordi Atserias	Yahoo! Research (Spain)
Maite Melero	Barcelona Media Innovation Center (Spain)
Martí Quixal	Barcelona Media Innovation Center (Spain)

Workshop Programme Committee

Toni Badia	Universitat Pompeu Fabra (Spain)
Rafael Banchs	Institute for Infocomm Research (Singapore)
Richard Beaufort	Université Catholique de Louvain(Belgium)
Steven Bedrick	Oregon Health & Science University
Louise-Amélie Cougnon	Université Catholique de Louvain(Belgium)
Jennifer Foster	Dublin City University (Ireland)
Michael Gamon	Microsoft Research (USA)
Dídac Hita	Infojobs (Spain)
Fei Liu	Bosch Research (USA)
Daniel Lopresti	Lehigh University (USA)
Ulrike Pado	VICO Research&Consulting GmbH
Lluís Padró	Universitat Politècnica de Catalunya (Spain)
Alan Ritter	CSE, University of Washington (USA)

Preface

The Web 2.0 has transferred the authorship of contents from institutions to the people; the web has become a channel where users exchange, explain or write about their lives and interests, give opinions and rate others' opinions. The so-called User Generated Content (UGC) in text form is a valuable resource that can be exploited for many purposes, such as cross-lingual information retrieval, opinion mining, enhanced web search, social science analysis, intelligent advertising, and so on.

In order to mine the data from the Web 2.0 we first need to understand its contents. Analysis of UG content is challenging because of its casual language, with plenty of abbreviations, slang, domain specific terms and, compared to published edited text, with a higher rate of spelling and grammar errors. Standard NLP techniques, which are used to analyze text and provide formal representations of surface data, have been typically developed to deal with standard language and may not yield the expected results on UGC. For example, shortened or misspelled words, which are very frequent in the Web 2.0 informal style, increase the variability in the forms for expressing a single concept.

This workshop aims at providing a meeting point for researchers working in the processing of UGC in textual form in one way or another, as well as developers of UGC-based applications and technologies, both from industry and academia.

Saturday 26 May

14:00 –16:00

Chairperson: Jordi Atserias

Comparing user generated content published in different social media sources

Óscar Muñoz-García and Carlos Navarro (Havas Media)

The growth of social media has populated the Web with valuable user generated content that can be exploited for many different and interesting purposes, such as, explaining or predicting real world outcomes through opinion mining. In this context, natural language processing techniques are a key technology for analysing user generated content. Such content is characterised by its casual language, with short texts, misspellings, and set-phrases, among other characteristics that challenge content analysis. This paper shows the differences of the language used in heterogeneous social media sources, by analysing the distribution of the part-of-speech categories extracted from the analysis of the morphology of a sample of texts published in such sources. In addition, we evaluate the performance of three natural language processing techniques (i.e., language identification, sentiment analysis, and topic identification) showing the differences on accuracy when applying such techniques to different types of user generated content.

Assigning part-of-speech to Dutch tweets

Mehdi Aminian, Tetske Avontuur, Zeynep Azar, Iris Balemans, Laura Elshof, Rose Newell, Nanne van Noord, Alexandros Ntavelos, Menno van Zaanen (Tilburg University)

In this article we describe the development of a part-of-speech (POS) tagger for Dutch messages from the Twitter microblogging website. Initially we developed a POS tag set ourselves with the intention of building a corresponding tagger from scratch. However, it turned out that the output of Frog, an existing high-quality POS tagger for Dutch, is of such quality that we decided to develop a conversion tool that modifies the output of Frog. The conversion consists of retokenization and adding Twitter-specific tags. Frog annotates Dutch texts with the extensive D-Coi POS tag set, which is used in several corpus annotation projects in the Netherlands. We evaluated the resulting automatic annotation against a manually annotated sub-set of tweets. The annotation of tweets in this sub-set has a high inter-annotator agreement and our extension of Frog shows an accuracy of around 95%. The add-on conversion tool that adds Twitter-specific tags to the output of Frog will be made available to other users.

Challenges in developing opinion mining tools for social media

Diana Maynard, Kalina Bontcheva, Dominic Rout (University of Sheffield)

While much work has recently focused on the analysis of social media in order to get a feel for what people think about current topics of interest, there are, however, still many challenges to be faced. Text mining systems originally designed for more regular kinds of texts such as news articles may need to be adapted to deal with facebook posts, tweets etc. In this paper, we discuss a variety of issues related to opinion mining from social media, and the challenges they impose on a Natural Language Processing (NLP) system, along with two example applications we have developed in very different domains. In contrast with the majority of opinion mining work which uses machine learning techniques, we have developed a modular rule-based approach which performs shallow linguistic analysis and builds on a number of linguistic subcomponents to generate the final opinion polarity and score.

A Qualitative Analysis of Informality Levels In Web 2.0 Texts: The Facebook Case Study

Alejandro Mosquera and Paloma Moreda (University of Alicante)

The study of the language used in Web 2.0 applications such as social networks, blogging platforms or on-line chats is a very interesting topic and can be used to test linguistic or social theories. However the existence of language deviations such as typos, emoticons, abuse of acronyms and domain-specific slang makes any linguistic analysis challenging. The characterization of this informal writing can be used to test the performance of Natural Language Processing tools when analysing Web 2.0 texts, where informality can play an important role. By being one of the most popular social media websites, Facebook handles an increasing volume of text, video and image data within its user profiles. In this paper, we aim to perform a qualitative analysis of informality levels in textual information publicly available on Facebook. In particular, this study focus on developing informality dimensions, a set of meaningful and comparable variables, discovered by mapping textual features by affinity and using unsupervised machine learning techniques. In addition, we explore the relation of informality and Facebook metadata such as received likes, gender, time range and publication type.

16:30 –18:00
Chairperson: Maite Melero

What is the text of a Tweet?

Joan Codina and Jordi Atserias (Fundació Barcelona Media)

Twitter is a popular micro blogging/social medium for broadcasting news, staying in touch with friends and sharing opinions using up to 140 characters per message. In general, User generated Content (e.g. Blogs, Tweets) differs from the kind of text the traditional Natural Language Processing tools have been developed and trained. The use of non-standard language, emoticons, spelling errors, letter casing, unusual punctuation makes applying NLP to user generated content still an open issue. This work will focus on the effect of the Twitter metalanguage elements in the text processing, specifically for PoS tagging. Several different strategies to deal with twitter specific metalanguage elements are presented and evaluated. The results show that it is necessary to remove metalanguage elements. However some text normalisation or PoS tagger adaptation is needed in order to have a clear evaluation about which of the different methods to treat twitter metalanguage elements is better.

Linguistic Resources for Genre-Independent Language Technologies: User-Generated Content in BOLT

Jennifer Garland, Stephanie Strassel, Safa Ismael, Zhiyi Song, Haejoong Lee (Linguistic Data Consortium, University of Pennsylvania)

We describe an ongoing effort to collect and annotate very large corpora of user-contributed content in multiple languages for the DARPA BOLT program, which has among its goals the development of genre-independent machine translation and information retrieval systems. Initial work includes collection of several hundred million words of online discussion forum threads in English, Chinese and Egyptian Arabic, with multi-layered linguistic annotation for a portion of the collected data. Future phases will target still more challenging genres like Twitter and text messaging. We provide details of the collection strategy and review some of the particular technical and annotation challenges stemming from these genres, and conclude with a discussion of strategies for tackling these issues.

Collaborative Resource Development and Delivery

27 May 2012

ABSTRACTS

Editors:

Nancy Ide, Collin Baker, Christiane Fellbaum, Rebecca Passonneau

Workshop Programme

09:00 –09:15 – Welcome and Overview

09:15 –10:00 – Discussion paper

The MASC/MultiMASC Community Collaboration Project: Why You Should Be Involved and How
Nancy Ide, Collin Baker, Christiane Fellbaum, Rebecca Passonneau

10:00 –10:30 – Discussion: Strategies to Engage the Community in Collaborative Annotation

10:30 –11:00 Coffee break

11:00 –11:30 – Invited talk

Towards a Linguistic Linked Open Data Cloud
Christian Chiarcos

11:30 –12:30 – Collaborative annotation task and discussion

12:30 –14:00 – Lunch break

14:00 –15:30 Paper session

Annotated Corpora in the Cloud: Free Storage and Free Delivery
Graham Wilcock

Guidance through the Standards Jungle for Linguistic Resources
Maik Stührenberg, Antonina Werthmann, Andreas Witt

Supporting Collaborative Improvement of Resources in the Khresmoi Health Information System
Lorraine Goeuriot, Allan Hanbury, Gareth J. F. Jones, Liadh Kelly, Sascha Kriewel, Ivan Martinez Rodriguez, Henning Müller, Miguel A. Tinte

15:30 –16:00 – Demonstrations

16:00 –16:30 Coffee break

16:30 –17:40 – Paper session

Building Parallel Corpora Through Social Network Gaming
Nathan David Green

Three Steps for Creating High-Quality Ontology Lexica
John McCrae, Philipp Cimiano

PromONTotion: Creating an Advertisement Thesaurus By Semantically Annotating Ad Videos Through Collaborative Gaming
Katia Lida Kermanidis, Emmanouil Maragkoudakis

The Phrase Detective Multilingual Corpus, Release 0.1
Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, Luca Ducceschi

17:40 –18:00 – Closing

Workshop Organizers

Nancy Ide	Vassar College, USA
Collin Baker	ICSI/UC Berkeley, USA
Christiane Fellbaum	Princeton Univeristy, USA
Rebecca Passonneau	Columbia University, USA

Workshop Programme Committee

Nicoletta Calzolari	ILC/CNR, Italy
Bob Carpenter	Alias I, Inc., USA
Chris Cieri	LDC, University of Pennsylvania, USA
Bill Dolan	Microsoft Corp., USA
Dan Flickinger	Stanford University, USA
Terry Langendoen	NSF and University of Arizona, USA
Massimo Poesio	University of Trento, Italy
Sameer Pradhan	BBN Technologies, USA
James Pustejovsky	Brandeis University, USA
Owen Rambow	Columbia University, USA
Manfred Stede	Universität Potsdam, Germany

Preface

A confluence of needs and activities points to a new emphasis in computational linguistics to address lexical, propositional, and discourse semantics through corpora. A few examples are:

- the demand for high quality linguistic annotations of corpora representing a wide range of phenomena, especially at the semantic level, to support machine learning and computational linguistics research in general;
- the demand for high quality annotated corpora representing a broad range of genres that are flexible and extensible as need demands;
- the demand for high quality lexical and semantic resources to incorporate into the annotation process, and for the annotation process to produce;
- the need for easy-to-use, open access to all of these resources for everyone.

Such resources can be very costly to produce, due to the need for manual creation or validation to ensure quality. Therefore, to answer the growing need and lower the costs of resource creation and enhancement, there is a movement within the community toward collaborative resource development, including collaborative corpus annotation and collective creation/enhancement of lexical resources and knowledge bases. Collaborative development encompasses both engaging the community in annotation and development of common resources, as well as crowdsourcing, gaming, and similar solutions. The papers in this workshop address both of these approaches to collaborative development, as well as software to support this development and other issues such as the role of standards for collaboratively created resources.

This workshop was motivated by a meeting of eminent researchers and developers of language resources, held at Columbia University in New York City in October, 2011. The goal of the meeting was to explore ways to involve the natural language processing community in the development of language resources—most notably, annotated linguistic corpora—in order to offset the high costs of resource creation. The focus of the discussions was the Manually Annotated Sub-Corpus (MASC) (http://www.anc.org/MASC), a community-based collaborative annotation project that is intended to provide the basis for development of a resource that is richly annotated for both variety and variants of linguistic phenomena. Among the conclusions of that workshop was a decision to broaden the discussion to include the community a whole by holding a workshop at LREC. This workshop therefore includes a special session devoted to strategies for engaging the community in collaborative linguistic annotation projects such as MASC.

The workshop also includes a collaborative annotation task that will engage all participants in the annotation of multiple phenomena over a common text. A following discussion session considers the results in order to address issues such as the level of agreement among the participants on the various tasks, and what it suggests in terms of the viability of collaborative annotation and crowdsourcing for creating high-quality linguistic annotations; and ways in which annotations on multiple levels may be used collectively to improve overall quality and contribute to analysis.

Discussion Paper

Sunday 27 May, 9:15 – 10:00
Chairperson: Massimo Poesio

The MASC/MultiMASC Community Collaboration Project: Why You Should Be Involved and How

Nancy Ide, Collin Baker, Christiane Fellbaum, Rebecca Passonneau

The Manually Annotated Sub-Corpus (MASC) Project has developed a half million word corpus with multiple levels of annotation, intended to serve as a basis for a community collaborative effort that would enhance the resource with contributed annotations and data. The MASC team converts all annotations, in-house or contributed, to a common format so they can be easily used together and compared. We are now in the process of launching a new effort to build corpora comparable to MASC in other languages, which will demand community involvement to not only annotate but also assemble the corpora. However, it is not clear how much community engagement we can expect, for although some MASC data and annotations have been freely available for over a year, we have received very few contributions of annotations. We outline here some possible reasons why there has not been more community engagement with MASC, and, more generally, consider whether or not it is realistic to rely on community contributions for resource development and enhancement. We also suggest several strategies for augmenting community involvement, to serve as a basis for a broader discussion.

Invited Paper

Sunday 27 May, 11:00 –11:30
Chairperson: Nancy Ide

Towards a Linguistic Linked Open Data Cloud

Christian Chiarcos

I describe benefits of modeling linguistic resources as Linked Data, i.e., using RDF, publishing them under an open license, and creating links between them. Further, an overview over currently on-going community efforts to create a Linked Open Data sub-cloud of linguistic resources will be given. Both aspects are illustrated for the MASC corpus.

Collaborative Annotation Task

Sunday 27 May, 11:30 –12:30
Chairpersons: Collin Baker, Christiane Fellbaum, Nancy Ide

The Collaborative Annotation task engages the workshop participants in annotating the same short text for different phenomena, including anaphora, semantic roles, and sense assignment. The text is automatically pre-tagged for part-of-speech and shallow parse. Results will be summarized in a later discussion session, in which issues such as the following will be addressed:

- How much agreement (or lack of agreement) is shown for the various phenomena? If we "experts" cannot agree, how much error are we willing to accept from the "crowd"?
- What is the relative difficulty of the different annotation types, and what does this suggest about expert vs. crowd vs. automatic annotation?
- How might annotations of the different phenomena be constructively used together for

annotation or analysis?
- To what extent do errors in the automatically-labeled phenomena get in the way of combining the annotations to produce a synthesis?
- What insights were gained during the annotation?

Sunday 27 May, 14:00 – 15:30
Chairperson: Collin Baker

Annotated Corpora in the Cloud: Free Storage and Free Delivery

Graham Wilcock

The paper describes a technical strategy for implementing natural language processing applications in the cloud. Annotated corpora can be stored in the cloud and queried in normal web browsers via user interfaces implemented in the described framework. A key aim of the strategy is to exploit the free storage and processing that is available in the cloud, while avoiding lock-in to proprietary infrastructure. A half-million-word annotated corpus application is described as a working example of the strategy.

Guidance through the standards jungle for linguistic resources

Maik Stührenberg, Antonina Werthmann, Andreas Witt

Research today is often performed in collaborated projects composed of project partners with different backgrounds and from different institutions and countries. Standards can be a crucial tool to help harmonizing these differences and to create sustainable resources. However, choosing a standard depends on having enough information to evaluate and compare different annotation and metadata formats. In this paper we present ongoing work on an interactive, collaborative website that collects information on standards in the field of linguistics as a means to guide interested researchers.

Supporting Collaborative Improvement of Resources in the Khresmoi Health Information System

Lorraine Goeuriot, Allan Hanbury Gareth J. F. Jones, Liadh Kelly, Sascha Kriewel, Ivan Martinez Rodriguez, Henning Müller, Miguel A. Tinte

Since medical knowledge relies on both scientific knowledge and real-life experience, the importance of user contributions to improve resources in health systems cannot be underestimated. We present work from the Khresmoi project, which aims to develop a multilingual multimodal search and access system for biomedical information and documents. Khresmoi targets three distinct user classes with differing levels of medical knowledge and information requirements, namely: general public, general practitioners, and, as an example of an area of clinical expertise, radiologists. The Khresmoi system will provide these users with valuable (whose quality has been evaluated and approved) and enriched (meta information from biomedical knowledge bases is added) medical information, selected to fit their medical knowledge and their preferred language. The system will include novel collaborative components of the system are designed to provide means for users to contribute to the system's knowledge by adding or correcting annotations to the documents, as well as a collaborative platform where they will be able to share their own files and both annotate and discuss them.

Sunday 27 May, 16:30 –17:40
Chairperson: Christiane Fellbaum

Building parallel corpora through social network gaming

Nathan David Green

Building training data is labor-intensive and presents a major obstacle to the advancement of Natural Language Processing (NLP) systems. A prime use of NLP technologies has been toward the construction machine translation systems. The most common form of machine translation systems are phrase based systems that require extensive training data. Building this training data is both expensive and error prone. Emerging technologies, such as social networks and serious games, offer a unique opportunity to change how we construct training data. These serious games, or games with a purpose, have been constructed for sentence segmentation, image labeling, and co-reference resolution. These games work on three levels: They provide entertainment to the players, the reinforce information the player might be learning, and they provide data to researchers. Most of these systems, while well intended and well developed, have lacked participation.
We present a set of linguistically based games that aim to construct parallel corpora for a multitude of languages and allow players to start learning and improving their own vocabulary in these languages. As of the first release of the games, GlobeOtter is available on Facebook as a social network game. The release of this game is meant to change the default position in the field, from creating games that only linguists play, to releasing linguistic games on a platform that has a natural user base and ability to grow.

Three steps for creating high-quality ontology lexica

John McCrae, Philipp Cimiano

Sophisticated NLP applications working on particular domains require rich information on both the linguistic properties of words and the semantics of these words. We propose a three-step methodology for the creation of high-quality ontology-lexica, which combine detailed syntactic information with deep semantic information about words and their associated meanings. Our proposed method consists of three steps: first we rely on a standard NLP pipeline to create a preliminary version of the ontology lexicon automatically. In this step, the automatically created lexicon is linked to existing legacy lexical resources. The second step involves referencing existing lexical and semantic resources and importing data. Finally, a manual review step is required that is supported by a novel editor to facilitate the inspection and manual validation and modification and thus continuous refinement and improvement of the ontology lexicon.

PromONTotion: Creating an Advertisement Thesaurus By Semantically Annotating Ad Videos Through Collaborative Gaming

Katia Lida Kermanidis, Emmanouil Maragkoudakis

The present work describes the plan of PromONTotion, a ready to launch research project that aims at creating a semantic thesaurus of the advertising domain. The resource will be developed collaboratively using crowdsourcing. A web-based game, entertaining enough to keep the player's interest active for a long time, will be designed for the collaborative semantic annotation of the content of ad videos. The inserted terms will populate the thesaurus, a hierarchical structure formed by concepts, concept attributes and semantic relations among them. Advertisers will access the thesaurus through a friendly interface, which will allow them to have full access to the capabilities of the resource. The ad videos, the terminology, statistical information regarding co occurrence of concepts and attributes, statistical information regarding the impact the ads had on the annotators-players will be available to the advertiser for supporting him in the creative process of designing a new ad campaign.

The Phrase Detective Multilingual Corpus, Release 0.1

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz, Livio Robaldo, Luca Ducceschi

The Phrase Detectives Game-With-A-Purpose for anaphoric annotation has been live since December 2008, collecting over 2.5 million judgments on the anaphoric expressions in texts in two languages (English and Italian) from around 9,000 players. In this paper we summarize our recent work on creating a corpus using these annotations.

Language Resources for Public Security Applications

http://www.lrps.amu.edu.pl/

27 May 2012

ABSTRACTS

Editors:

Zygmunt Vetulani, Edouard Geoffrois

Technical editor: Wojciech Czarnecki

Workshop Programme

14:00 – 14:20 – Opening and introductory presentation by Zygmunt Vetulani and Edouard Geoffrois

Zygmunt Vetulani, Edouard Geoffrois, Wojciech Czarnecki and Bartłomiej Kochanowski, *Language Resources for Public Security Applications: Needs and Specificities*

14:20 – 15:00 – Invited Keynote Talk by Chris Cieri (University of Pennsylvania, USA, Linguistic Data Consortium, Executive Director) , *Language Resources for Public Security Applications: a Data Center Perspective*

15:00 – 16:00 – Resources (oral presentations)

Adam Dąbrowski, Szymon Drgas, Paweł Pawłowski and Julian Balcerek, *Development of PUEPS - corpus of emergency telephone conversations*

Irina Temnikova and K. Bretonnel Cohen, *The Crisis Management Corpus and its Application to the Study of the Crisis Management Sub-language*

Christian Fluhr, Aurélie Rossi, Louise Boucheseche and Fadhela Kerdjoudj, *Extraction of information on activities of persons suspected of illegal activities from web open sources*

16:00 – 16:30 Coffee break

16:30 – 17:15 – Applications (poster presentations)

Carlo Aliprandi, Tomas By and Sérgio Paulo, *Language Processing and Linguistic Data in the CAPER Project*

Richard Beaufort, Alexander Panchenko and Cédrick Fairon, *Detection of Child Sexual Abuse Media on P2P Networks: Normalization and Classification of Associated Filenames*

Simona Cantarella, Carlo Ferigato and Evans Boateng Owusu, *Design of a Controlled Language for Critical Infrastructures Protection*

Ales Horak, Karel Pala and Jan Rygl, *Authorship Identification to Improve Public Security*

Wiesław Lubaszewski and Michał Korzycki, *Unexpected Factual Associations Mining*

Miriam R L Petruck and Gerard de Melo, *Precedes: A Semantic Relation in FrameNet*

Milan Rusko, Sakhia Darjaa, Marian Trnka, Miloš Cerňak, *Expressive speech synthesis database for emergent messages and warnings generation in critical situations*

Zygmunt Vetulani, *Language Resources in a Public Security Application with Text Understanding Competence. A Case Study: POLINT-112-SMS*

17:15 – 18:00 – General discussion (animated by Frédérique Segond)

Workshop Organizers

Zygmunt Vetulani	Adam Mickiewicz University in Poznań, Poland
Edouard Geoffrois	Direction Générale de l'Armement, Mission for Scientific Research and Innovation (MRIS), Bagneux, France

Workshop Programme Committee

Laura Chaubard	Direction Générale de l'Armement, DGA Ingénierie de Projets, Bagneux, France
Edouard Geoffrois	Direction Générale de l'Armement, Mission for Scientific Research and Innovation (MRIS), Bagneux, France
Jakub Gorczyński	Polish National Police, Poznań, Poland
Fryni Kakoyianni	University of Cyprus, Nicosia, Cyprus
Nasrullah Memon	University of Southern Denmark, Odense, Denmark
Mario Montoleone	Salerno University, Italy
Karel Pala	Masaryk University, Brno, Czech Republic
Frédérique Segond	Viseo, Grenoble, France
Tadeusz Tomaszewski	University of Warsaw, Poland
Zygmunt Vetulani	Adam Mickiewicz University in Poznań, Poland

Preface/Introduction

Workshop Description

Public security in Europe and in the World is facing several threats. These include threats connected with intended human activities such as terrorism, spontaneous risks related to uncontrolled behavior of individuals involved in mass events, natural disasters, etc. Combating these dangers generates challenges for information and communication technologies which in many cases directly involve various forms of natural language processing. Gathering, maintaining and processing language resources specific for security applications is of primary importance for the language technologies concerned. In some cases it appears useful to investigate and use sensitive linguistic data which generates technological and legal problems connected with privacy, ownership, civic rightsprotection, etc.

The workshop is intended to serve as a thematic discussion forum open to:

- language resources suppliers,
- researchers and language engineers interested in the development of systems for security applications involving language technologies,
- potential/actual users of such systems,
- people concerned with legal aspects of gathering, maintenance and applications of language resources for public security purposes.

Generation of a long term cooperation projects involving the workshop participants would be adesired side effect of the workshop.

Areas of Interest

The workshop will focus on the knowledge serving applications serving public security. Particularemphasis will be given to the crucial role of language resources and related technologies. Contributions are invited on – but not limited to – the following topics:

- security specific corpora,
- security specific terminology,
- language models for specific sub-languages and language registers important for security research,
- language technology based tools to enhance public security,
- linguistic tools for risk assessment,
- controlled languages for public security applications,
- AI and NLP decision supporting systems,
- sharing and processing sensitive linguistic data,
- legal aspects of security-oriented natural language processing and engineering,
- access to sensitive data,
- IPR issues,
- protection and use of sensitive source data,
- international collaboration issues,
- issues related with national and international funding

Opening and introductory presentation

Sunday 27 May, 14:00 – 14:20
Chairperson: Zygmunt Vetulani

Language Resources for Public Security Applications: Needs and Specificities

Zygmunt Vetulani, Edouard Geoffrois, Wojciech Czarnecki and Bartłomiej Kochanowski

Language technologies and the associated language resources necessary to develop them are needed in a number of applications in the public security sector, and there is a growing demand for such applications. The paper illustrates the scope and importance of the needs by presenting various examples of applications along with the corresponding language technologies and language resources. However, collecting and sharing these resources can be especially difficult in that sector due to its specificities. The paper proposes to better identify and acknowledge these specificities in order to better address them and suggests that sharing experience across the various applications within the sector might help to overcome the difficulties.

Invited Keynote Lecture

Sunday 27 May, 14:20 – 15:00
Chairperson: Edouard Geoffrois

Language Resources for Public Security Applications: a Data Center Perspective

Chris Cieri

Among the many corpora that LDC is producing or distributing, several, for example some of the Mixer corpora, are related to public security variously defined. In this talk we present some of these corpora and how they were created. We also describe some of the issues encountered in their creation which are related to the public security domain, how we overcame them and the lessons learned. Some specific issues we will discuss include matching data specifications to rapidly evolving requirements, managing intellectual property, protecting the privacy of human subjects and distributing resulting data.

Resources (oral presentations)

Sunday 27 May, 15:00 – 16:00
Chairperson: Edouard Geoffrois

Development of PUEPS - corpus of emergency telephone conversations

Adam Dąbrowski, Szymon Drgas, Paweł Pawłowski and Julian Balcerek

In this article development of a PUEPS corpus is described. This dataset contains recordings of the acted emergency telephone conversations. Speakers that participated in the experiments reported crime scenes that were presented to them in a form the earlier prepared movies. Recording sessions were performed in the laboratory conditions. To each conversation metadata that summarize information about the speaker, conversation, and the reported event were added. Moreover, manually prepared transcriptions enriched with tags describing paralinguistic phenomena are also a part of the described corpus. These transcriptions were made using tools prepared by the authors for fast and convenient work due to: prompting, annotation, and data management mechanisms. The transcription experiments showed substantial improvement of the work efficiency and speed. Final multilevel speaker recognition experiments proved that the accuracy of the speaker recognition is noticeably improved due to the use of transcriptions and the linguistic level analysis.

The Crisis Management Corpus and its Application to the Study of the Crisis Management Sub-language

Irina Temnikova and K. Bretonnel Cohen

The Crisis Management Corpus and its Application to the Study of the Crisis Management Sub-language Irina Temnikova and K. Bretonnel Cohen This article presents a novel language resource, the Crisis Management Corpus (CMC). The corpus is the first in its domain and is expected to be of utility for linguistic studies and for natural language processing applications in the crisis management and the public security domains. The article describes the collection, pre-processing and composition of this resource, along with its possible applications. Two example applications of the resource are described in detail. The first application is the study of the text complexity levels characterizing the CMC, with the aim of evaluating the communicative efficiency of written documents in the domain. The second application is a preliminary investigation of the linguistic characteristics of the crisis management sub-language.

Extraction of information on activities of persons suspected of illegal activities from web open sources

Christian Fluhr, Aurélie Rossi, Louise Boucheseche and Fadhela Kerdjoudj

This work is part of the French funded SAIMSI project (Suivi Adaptatif Interlingue et Multisource des Informations). The aim of the project is to follow activities of persons suspected of illegal actions like terrorism, drug traffic or money laundering. The paper specially focuses on the information extraction. This extraction is done in French, English, Arabic and Chinese. The information extraction is based on a deep morphosyntactic analysis. Recognition of single words, idiomatic expressions, compounds is performed and named entities are identified and categorized. Dependency relations are built, passive/active forms, negation anaphora, verb tenses are processed. Information extraction is application-independent and uses extraction rules. At this level some named entity categories can be reconsidered. This extraction is based on a large ontology of the security. The paper details the problems of the consolidation of the extracted knowledge at the document level. The future evaluation on WEPS-3 data is presented.

Applications (poster presentations)

Sunday 27 May, 16:30 – 17:15
Chairperson: Zygmunt Vetulani

Language Processing and Linguistic Data in the CAPER Project

Carlo Aliprandi, Tomas By and Sérgio Paulo

Much information of potential relevance to police investigations of organised crime is available in public sources without being recognised and used. Barriers to the simple and efficient exploitation of this information include that not everything is easily searchable, and may be written in a language other than that of the investigator. To help overcome these problems, the CAPER project aims to create an integrated platform for acquisition, processing, and analysis of information in multiple languages, and also link this to legacy police IT systems. Full Natural Language Processing pipelines for multiple languages and media are used to map persons and organisations to actions and events, and Multi-lingual lexicons and gazetteers allow cross-lingual search in the indexed data. Domain-specific lexicons contain words and slang expressions with special senses in the context of organised crime. The system supports multilingual analysis of unstructured and audiovisual contents, based on text mining for fourteen languages, and uses language-neutral interfaces, so that addition of further languages will not require any modification of existing components.

Detection of Child Sexual Abuse Media on P2P Networks: Normalization and Classification of Associated Filenames

Richard Beaufort, Alexander Panchenko and Cédrick Fairon

The goal of the iCOP project is to build a system detecting the originators of pedophile content on P2P networks such as BitTorrent, eDonkey, or Kad. This paper outlines the key functions of the language processing in the iCOP system. Next, we describe the architecture of the language analysis module and its key components - filename classifier, term extractor, and filename normalizer. The language resources used in each component are discussed. The paper is also presenting the first experiments with the module on the standard porn data (used in the preliminary tests as a substitute of child pornography data). Our results show that the module is able to separate titles of the pornographic galleries and videos from the titles of encyclopaedia articles with accuracy up to 97%. Finally, we discuss the directions for the future research and developments of the iCOP language analysis module.

Design of a Controlled Language for Critical Infrastructures Protection

Simona Cantarella, Carlo Ferigato and Evans Boateng Owusu

We describe a project for the construction of controlled language for critical infrastructures protection} (CIP). This project originates from the need to coordinate and categorize the communications on CIP at the European level. These communications can be physically represented by official documents, reports on incidents, informal communications and plain e-mail. We explore the application of traditional library science tools for the construction of controlled languages in order to achieve our goal. Our starting point is an analogous work done during the sixties in the field of nuclear science known as the Euratom Thesaurus.

Authorship Identification to Improve Public Security

Ales Horak, Karel Pala and Jan Rygl

In the paper, we present details of a new project aimed at automatic web document analysis for the purpose of authorship attribution based on various stylistic and grammatical features of the text. We describe the corresponding system modules with their expected functionality and provide examples of text processing and evaluating techniques.

Unexpected Factual Associations Mining

Wiesław Lubaszewski and Michał Korzycki

The paper describes the LSA (Latent Semantic Analysis) algorithm as a tool for mining unexpected factual associations from text corpora. Due to the fact that LSA performs well on text corpora built from short texts it can be a useful tool to analyse e-mails stored in the mail box, chats logs or Internet fora content. Therefore the LSA may serve as a tool in forensic or security analysis.

Precedes: A Semantic Relation in FrameNet

Miriam R L Petruck and Gerard de Melo

Precedes: A Semantic Relation in FrameNet Miriam R. L. Petruck and Gerard de Melo International Computer Science Institute Berkeley, California, USA miriamp@icsi.berkeley.edu, demelo@icsi.berkeley.edu Abstract Automatic language processing systems depend on, among others factors, the effectiveness in modeling human cognitive abilities, including the capacity to draw inferences about prototypical or expected sequences of events and their temporal order. Appropriate response to a crisis is as important for public security as are efforts to prevent any such natural or man made disaster. Recent research (Mehrota et al. 2008) has recognized the need for accurate and actionable situation awareness during emergencies, where timely status updates are critical for effective crisis management. The present paper constitutes a contribution to situation awareness for Natural Language Processing (NLP) applications to improve communication among first responders, and features the frame-to-frame semantic relation Precedes, as implemented in FrameNet (http://framenet.icsi.berkeley.edu). Specifically, this work demonstrates the necessity and importance of the information encoded with Precedes for NLP applications, advocating the inclusion of such information in systems for security applications.

Expressive speech synthesis database for emergent messages and warnings generation in critical situations

Milan Rusko, Sakhia Darjaa, Marian Trnka, Miloš Cerňak

Automatic information and warning systems can be used to inform, warn, instruct and navigate people in dangerous and critical situations, and increase the effectiveness of crisis management and rescue operations. One of the activities in the frame of the EU SF project CRISIS is called “Extremely expressive (hyper-expressive) speech synthesis for urgent warning messages generation”. It is aimed at research and development enabling the possibility to design speech synthesizers with high naturalness and intelligibility in Slovak which will be capable of generating messages with various expressive loads. The synthesizer will be applicable to generate warning system messages in case of fire, flood, state security threats, etc. Early warning in relation to the above can be made thanks to fire and flood spread forecasting; modeling thereof is covered by other activities of the CRISIS project. The most important part needed for synthesizer building is the speech database. A method is proposed to create such a database. The first version of the expressive speech database is introduced and first experiments with expressive synthesizers trained with this database are discussed.

Language Resources in a Public Security Application with Text Understanding Competence. A Case Study: POLINT-112-SMS

Zygmunt Vetulani

The aim of this paper is to show the importance of language resources in the development of complex, public security oriented applications with natural language understanding components as essential parts of the system. We present a case study of a mature project in the public security sector. This case study aims at giving an idea of the spectrum of needs and problems, without pretention to exhaust the topic. As it is typical for public security oriented projects, besides usual problems due to the gaps in available language data (resources), designers and developers of the presented system needed to deal with sensible data necessary for efficient language modeling. To make the paper self-contained, we start with a compact presentation of the POLINT-112-SMS system. Then we present the language resources we used.

5th Workshop on the Representation and Processing of Sign Languages: Interactions between Corpus and Lexicon

27 May 2012

ABSTRACTS

Editors:

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

Workshop Programme

09:00 – 10:30	Oral Session A: *Relations between Corpus and Lexicon*
10:30 – 11:00	Coffee break
11:00 – 13:00	Poster Session B: *Methodology and Technology*
13:00 – 14:00	Lunch break
14:00 – 16:00	Poster Session C: *Resources*
16:00 – 16:30	Coffee break
16:30 – 19:00	Oral Session D: *Issues in the Construction of Sign Corpora*

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
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
Helen Cooper	University of Surrey, Guildford GB
Kearsy Cormier	Deafness Cognition and Language Research Centre, London GB
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
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
Matt Huenerfauth	City University of New York, New York 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, Athens GR
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
Meike Vaupel	University of Applied Sciences Zwickau, Zwickau DE

Session A: Relations between Corpus and Lexicon

Sunday 27 May, 09:00 – 10:30
Chairperson: Onno Crasborn Oral Session

From Meaning to Signs and Back: Lexicography and the Swedish Sign Language Corpus

Johanna Mesch and Lars Wallin

In this paper, we will present the advantages of having a reference dictionary, and how having a corpus makes dictionary making easier and more effective. It also gives a new perspective on sign entries in the dictionary, for example, if a sign uses one or two hands, or which meaning 'genuine signs' have, and it helps find a model for categorization of polysynthetic signs that is not found in the dictionary. Categorizing glosses in the corpus work has compelled us to revisit the dictionary to add signs from the corpus that are not already in the dictionary and to improve sign entries already in the dictionary based on insights that have been gained while working on the corpus.

Linking an ID-gloss Database of ASL with Child Language Corpora

Julia Fanghella, Leah Geer, Jonathan Henner, Julie Hochgesang, Diane Lillo-Martin, Gaurav Mathur, Gene Mirus and Pedro Pascual-Villanueva

We describe an on-going project to develop a lexical database of American Sign Language (ASL) as a tool for annotating ASL corpora collected in the United States. Labs within our team complete locally chosen fields using their notation system of choice, and pick from globally available, agreed-upon fields, which are then merged into the global database. Here, we compare glosses in the database to annotations of spontaneous child data from the BiBiBi project (Chen Pichler et al., 2010). These comparisons validate our need to develop a digital link between the database and corpus. This link will help ensure that annotators use the appropriate ID-glosses and allow needed glosses to be readily detected (Johnston, 2011b; Hanke and Storz, 2008). An ID-gloss database is essential for consistent, systematic annotation of sign language corpora, as (Johnston, 2011b) has pointed out. Next steps in expanding and strengthening our database's connection to ASL corpora include (i) looking more carefully at the source of data (e.g. who is signing, language background, age, region, etc.), (ii) taking into account signing genre (e.g. presentation, informal conversation, child-directed etc.), and (iii) confronting the matter of deixis, gesture, depicting verbs and other constructions that depend on signing space.

Integrating Corpora and Dictionaries: Problems and Perspectives, with Particular Respect to the Treatment of Sign Language

Jette H. Kristoffersen and Thomas Troelsgård

In this paper, we will discuss different possibilities for integration of corpus data with dictionary data, mainly seen from a lexicographic point of view and in a sign language context. For about 25 years a text corpus has been considered a useful, if not necessary tool for editing dictionaries of written and spoken languages. Corpora are equally useful to sign language lexicographers, but sign language corpora have not become accessible until recent years. Nowadays corpora exist, or are being developed, for several sign languages, and we have the possibility of editing new, truly corpus-based sign language dictionaries, and of developing interfaces that integrate corpus and dictionary data. After a brief look at three existing integrated interfaces, one for German, one for Danish, and one for Danish Sign Language, we point out some of the problems that should be considered when making an integrated interface, and, finally, we briefly outline the future perspectives of integrated sign language corpus-dictionary interfaces.

Session B: Methodology and Technology

Sunday 27 May, 11:00 – 13:00
Chairperson: Johanna Mesch Poster Session

From Corpus to Lexical Database to Online Dictionary: Issues in Annotation of the BSL Corpus and the Development of BSL SignBank

Kearsy Cormier, Jordan Fenlon, Trevor Johnston, Ramas Rentelis, Adam Schembri, Katherine Rowley, Robert Adam and Bencie Woll

One requirement of a sign language corpus is that it should be machine-readable, but only a systematic approach to annotation that involves lemmatisation of the sign language glosses can make this possible at the present time. Such lemmatisation involves grouping morphological and phonological variants together into a single lemma, so that all related variants of a sign can be identified and analysed as a single sign. This lemmatisation process is made more straightforward by the existence of a comprehensive lexical database, as in the case for Australian Sign Language (Auslan). When annotation of data collected as part of the British Sign Language (BSL) Corpus Project began, no such lexical database for BSL existed. Therefore, a lemmatised BSL lexical database was created concurrently during annotation of the BSL Corpus data. As part of ongoing work by the Deafness Cognition & Language Research Centre, this lexical database is being developed into an online BSL dictionary, BSL SignBank. This paper describes the adaptation of the Auslan lexical database into a BSL lexical database, and the current development of this lexical database into BSL SignBank.

Linking Corpus NGT Annotations to a Lexical Database Using Open Source Tools ELAN and LEXUS

Onno Crasborn, Micha Hulsbosch and Han Sloetjes

This paper describes how we have made a first start with expanding the functionality of the ELAN annotation tool to create a bridge to a lexical database. A first lookup facility of an annotation in a LEXUS database is created, which generates a user-configurable selection of fields from that database, to be displayed in ELAN. In addition, an extension of the (open) controlled vocabularies that can be specified for tiers allows for the creation of very large vocabularies, such as lexical items in a language. Such an ‘external controlled vocabulary’ is an XML file that can be published on any web server and thus will be accessible to any interested party. Future development should allow for the vocabulary to be directly linked to a LEXUS database and thus also allow for access right management.

Improvements on the Distributed Architecture for Assisted Annotation of Video Corpora

Rémi Dubot and Christophe Collet

Progress on automatic annotation looks attractive for the research on sign languages. Unfortunately, such tools are not easy to deploy or share. We propose a solution to uncouple the annotation software from the automatic processing module.

Such a solution requires many developments: design of a network stack supporting the architecture, production of a video server handling trust policies, standardization of annotation encoding.

In this article, we detail the choices made to implement this architecture.

Semi-Automatic Annotation of Semantic Relations in a Swiss German Sign Language Lexicon

Sarah Ebling, Katja Tissi and Martin Volk

We propose an approach to semi-automatically obtaining semantic relations in Swiss German Sign Language (Deutschschweizerische Gebärdensprache, DSGS). We use a set of keywords including the gloss to represent each sign. We apply GermaNet, a lexicographic reference database for German annotated with semantic relations. The results show that approximately 60% of the semantic relations found for the German keywords associated with 9000 entries of a DSGS lexicon also apply for DSGS. We use the semantic relations to extract sub-types of the same type within the concept of double glossing (Konrad 2011). We were able to extract 53 sub-type pairs.

Sign Language Technologies and Resources of the Dicta-Sign Project

Eleni Efthimiou, Stavroula-Evita Fotinea, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos and François Lefebvre-Albaret

Here we present the outcomes of Dicta-Sign FP7-ICT project. Dicta-Sign researched ways to enable communication between Deaf individuals through the development of human-computer interfaces (HCI) for Deaf users, by means of Sign Language. It has researched and developed recognition and synthesis engines for sign languages (SLs) that have brought sign recognition and generation technologies significantly closer to authentic signing. In this context, Dicta-Sign has developed several technologies demonstrated via a sign language aware Web 2.0, combining work from the fields of sign language recognition, sign language animation via avatars and sign language resources and language models development, with the goal of allowing Deaf users to make, edit, and review avatar-based sign language contributions online, similar to the way people nowadays make text-based contributions on the Web.

Towards Tagging of Multi-Sign Lexemes and other Multi-Unit Structures

Thomas Hanke, Susanne König, Reiner Konrad and Gabriele Langer

With the building of larger sign language corpora tagging, handling and analysing large amounts of data reach a new level of complexity. Efficiency and interpersonal consistency in tagging are relevant issues as well as procedures and structures to identify and tag relevant linguistic units and structures beyond and above the manual sign level. We present and discuss problems and possible solution approaches (focussing on the working environment of iLex) of how to deal with multi-unit structures and more specifically multi-sign lexemes in annotation and lexicon building.

From Form to Function. A Database Approach to Handle Lexicon Building and Spotting Token Forms in Sign Languages

Reiner Konrad, Thomas Hanke, Susanne König, Gabriele Langer, Silke Matthes, Rie Nishio and Anja Regen

Using a database with type entries that are linked to token tags in transcripts has the advantage that consistency in lemmatising is not depending on ID-glosses. In iLex types are organised in different levels. The type hierarchy allows for analysing form, iconic value, and conventionalised meanings of a sign (sub-types). Tokens can be linked either to types or sub-types.

We expanded this structure for modelling sign inflection and modification as well as phonological variation. Differences between token and type form are grouped by features, called qualifiers, and specified by feature values (vocabularies). Built-in qualifiers allow for spotting the form difference when lemmatising. This facilitates lemma revision and helps to get a clear picture of how inflection, modification, or phonological variation is distributed among lexical signs. This is also a strong indicator for further POS tagging. In the long term this approach will extend the lexical database from citation-form closer to full-form.

The paper will explain the type hierarchy and introduce the qualifiers used up-to-date. Further on the handling and how the data are displayed will be illustrated. As we report work in progress in the context of the DGS corpus project, the modelling is far from complete.

A Conceptual Approach in Sign Language Classification for Concepts Network

Cedric Moreau

Most websites presuppose a conceptual equivalence between a written word and a sign. In such tools, signs which do not have strict written equivalent lexicons cannot be found. The collaborative website OCELLES project LSF/French tries to give the opportunity to obtain several signs for a unique concept, with the possibility of uploading a sign without being constrained by written language. Although word checking in a written text is quite easy, this is not the case for sign checking in a video. Today studies are carried out in the field of gesture recognition, but all the sign language linguistic parameters cannot be considered as such. Indeed, they have to be used simultaneously during communication interactions. Our approach based upon the semiological Cuxac model (Cuxac, 2000) and Thom morphogenesis theory (Thom, 1973), could help to find a sign in a sign dictionary without using any written language.

A New Web Interface to Facilitate Access to Corpora: Development of the ASLLRP Data Access Interface

Carol Neidle and Christian Vogler

A significant obstacle to broad utilization of corpora is the difficulty in gaining access to the specific subsets of data and annotations that may be relevant for particular types of research. With that in mind, we have developed a web-based Data Access Interface (DAI), to provide access to the expanding datasets of the American Sign Language Linguistic Research Project (ASLLRP). The DAI facilitates browsing the corpora, viewing videos and annotations, searching for phenomena of interest, and downloading selected materials from the website. The web interface, compared to providing videos and annotation files off-line, also greatly increases access by people that have no prior experience in working with linguistic annotation tools, and it opens the door to integrating the data with third-party applications on the desktop and in the mobile space. In this paper we give an overview of the available videos, annotations, and search functionality of the DAI, as well as plans for future enhancements. We also summarize best practices and key lessons learned that are crucial to the success of similar projects.

A Proposal for Making Corpora More Accessible for Synthesis: A Case Study Involving Pointing and Agreement Verbs

Rosalee Wolfe, John C. McDonald, Jorge Toro and Jerry Schnepp

Sign language corpora serve many purposes, including linguistic analysis, curation of endangered languages, and evaluation of linguistic theories. They also have the potential to serve as an invaluable resource for improving sign language synthesis. Making corpora more accessible for synthesis requires geometric as well as linguistic data. We explore alternate approaches and analyze the tradeoffs for the case of synthesizing indexing and agreement verbs. We conclude with a series of questions exploring the feasibility of utilizing corpora for synthesis.

Session C: Resources

Sunday 27 May, 14:00 – 16:00

Chairperson: Jette Kristoffersen Poster Session

SIGNSPEAK Project Tools: A Way to Improve the Communication Bridge between Signer and Hearing Communities

Javier Caminero, Mari Carmen Rodriguez-Gancedo, Alvaro Hernandez-Trapote and Beatriz Lopez-Mencia

The SIGNSPEAK project is aimed at developing a novel scientific approach for improving the communication between signer and hearing communities. In this way, SIGNSPEAK technology captures the video information from the signer and converts it into text. To do that, SIGNSPEAK consortium has devoted great efforts to the creation and annotation of the RWTH-Phoenix corpus. Based on it, a multimodal processing of the captured video is carried out and the resultant sign sequence is translated into natural language. Afterwards, the intended message could be communicated to hearing-able people using a text-to-speech (TTS) engine. In the reverse way, speech from hearing-able people would be transformed into text using Automatic Speech Recognition (ASR) and then the text would be processed by virtual avatars able to compose the suitable sign sequence. In SIGNSPEAK project, scientific and usability approaches have been combined to go beyond the state-of-the-art and contributing to suppress barriers between signer and hearing communities. In this work, a special stress was put in the development of a prototype and also, in setting of the grounds for future real industrial applications.

From Corpus to Lexicon: The Creation of ID-Glosses for the Corpus NGT

Onno Crasborn and Anne de Meijer

When glossing of the Corpus NGT started in 2007, there was no lexicon at our disposal to base ID-glosses on. Semantic labels were used without ensuring a constant relationship between sign form and gloss. This is currently being repaired by creating a lexicon from scratch alongside with the creation of new annotations. This substantial task is still in progress, but promises to lead to several new research avenues for the future. The current paper describes some of the choices that were made in the process, and specifies some of the glossing conventions that were used.

A GSL Continuous Phrase Corpus: Design and Acquisition

Athanasia-Lida Dimou, Vassilis Pitsikalis, Theodoros Goulas, Stavros Theodorakis, Panagiotis Karioris, Michalis Pissaris, Stavroula-Evita Fotinea, Eleni Efthimiou and Petros Maragos

The corpus presented in this article is composed of a limited number of Greek Sign Language (GSL) sentences and was created in order to provide additional data to the already obtained corpus during the first year of the Dicta-Sign project (Matthes et al., 2010). More specifically this corpus intended to serve as the ground upon which a significant part of the recognition process would be tested and evaluated, more precisely, the continuous sign language recognition algorithms developed in the project.

Given the targeted nature of this corpus we present here the constraints as well as the procedure followed in order to obtain it.

The procedure followed for the creation of this corpus, consists of its linguistic design and validation, the studio and hardware acquisition configuration, the implementation and supervision of the acquisition itself and the post-processing and annotation of the obtained data in order to release the set of usable annotated resources. The specific GSL phrase corpus forms the basis for machine learning and training to serve experimentation in the domain of continuous sign language processing and recognition.

A Study on Qualification/Naming Structures in Sign Languages

Michael Filhol and Annelies Braffort

In the prospect of animating virtual signers, this article addresses the issue of representing Sign, in particular on levels not restricted to the language lexicon. In order to choose and design a suitable model, we illustrate the main steps of our corpus-based methodology for linguistic structure identification and formal description with the example of a specific structure we have named "qualification/naming". We also discuss its similarity and difference with other Sign properties described in the literature such as compound signs. Consequently we explain our choice for a description model that does not separate lexicon and grammar in two disjoint levels for virtual signer input.

Experiences Collecting Motion Capture Data on Continuous Signing

Tommi Jantunen, Birgitta Burger, Danny De Weerdt, Irja Seilola and Tuija Wainio

This paper describes some of the experiences the authors have had collecting continuous motion capture data on Finnish Sign Language in the motion capture laboratory of the Department of Music at the University of Jyväskylä, Finland. Monologue and dialogue data have been recorded with an eight-camera optical motion capture system by tracking, at a frame rate of 120 Hz, the three-dimensional locations of small ball-shaped reflective markers attached to the signer's hands, arms, head, and torso. The main question from the point of view of data recording concerns marker placement, while the main themes discussed concerning data processing include gap-filling (i.e. the process of interpolating the information of missing frames on the basis of surrounding frames) and the importing of data into ELAN for subsequent segmentation (e.g. into signs and sentences). The paper will also demonstrate how the authors have analyzed the continuous motion capture data from the kinematic perspective.

Towards Russian Sign Language Synthesizer: Lexical Level

Alexey Karpov and Miloš Železný

In this paper, we present a survey of existing Russian sign language electronic and printed resources and dictionaries. The problem of differences in dialects of Russian sign language used in various local communities of Russia and some other CIS countries is discussed in the paper. Also the first version of a computer system for synthesis of elements of Russian sign language (signed Russian and fingerspelling) is presented in the given paper. It is a universal multi-modal synthesizer both for Russian spoken language and signed Russian that is based on a model of animated 3D signing avatar. The proposed system inputs data in the text form and converts them into the audio-visual modality, synchronizing visual manual gestures and articulation with audio speech signal. Generated audio-visual signed Russian speech and spoken language is a fusion of dynamic gestures shown by the avatar's both hands, lip movements articulating words and auditory speech, so the multimodal output is available both for the deaf and hearing-able people.

A Colorful First Glance at Data on Regional Variation Extracted from the DGS-Corpus: With a Focus on Procedures

Gabriele Langer

In this work in progress procedures for analyzing and displaying distributional patterns of sign variants have been developed and tested on data for color signs elicited by the DGS Corpus Project. The data for this preliminary study were elicited as isolated signs and have been made accessible through spot annotations in iLex. The annotations had not been lemma revised but nevertheless revealed some interesting insights. Several color signs exhibited a high degree of variation. The distributional maps showed that a number of signs were mainly used in certain regions and thus provided evidence on dialectal differences within DGS. The relevant information necessary to generate distributional maps have been directly extracted via SQL-statements from the corpus and fed into R. The approach is data driven. The distributional maps show either the distribution of one sign form (variant) or of several different variants in relation to each other. Analyses of regional distribution as displayed by the distributional maps may support the annotation and lemma revision process and are a valuable basis for a lexicographical description of signs and their use as needed for compiling dictionary entries. A refined procedure to take multiple regional influences on informants into account for analysis is proposed.

CUNY American Sign Language Motion-Capture Corpus: First Release

Pengfei Lu and Matt Huenerfauth

We are in the middle of a 5-year study to collect, annotate, and analyze an ASL motion-capture corpus of multi-sentential discourse. Now we are ready to release to the research community the first sub-portion of our corpus that has been checked for quality. This paper describes the recording and annotation procedure of our released corpus to enable researchers to determine if it would benefit their work. A focus of the collection process was the identification and use of prompting strategies for eliciting single-signer multi-sentential ASL discourse that maximizes the use of pronominal spatial reference yet minimizes the use of classifier predicates. The annotation of the corpus includes details about the establishment and use of pronominal spatial reference points in space. Using this data, we are seeking computational models of the referential use of signing space and of spatially inflected verb forms for use in American Sign Language (ASL) animations, which have accessibility applications for deaf users.

Dicta-Sign – Building a Multilingual Sign Language Corpus

Silke Matthes, Thomas Hanke, Anja Regen, Jakob Storz, Satu Worseck, Eleni Efthimiou, Athanasia-Lida Dimou, Annelies Braffort, John Glauert and Eva Safar

This paper presents the multilingual corpus of four European sign languages compiled in the framework of the Dicta-Sign project. Dicta-Sign researched ways to enable communication between Deaf individuals through the development of human-computer interfaces (HCI) for Deaf users, by means of sign language. Sign language resources were compiled to inform progress in the other research areas within the project, especially video recognition of signs, sign-to-sign translation, linguistic modelling, and sign generation. The aim for the corpus data collection was to achieve as high a level of naturalness as possible with semi-spontaneous utterances under lab conditions. At the same time the elicited data were supposed to be semantically close enough to be comparable both across individual informants and for all four sign languages. The sign language data were annotated using iLex and are now made available via a web portal that allows for different access options to the data.

Sign Language Resources in Sweden: Dictionary and Corpus

Johanna Mesch, Lars Wallin and Thomas Björkstrand

Sign language resources are necessary tools for adequately serving the needs of learners, teachers and researchers of signed languages. Among these resources, the Swedish Sign Language Dictionary was begun in 2008 and has been in development ever since. Today, it has approximately 8,000 sign entries. The Swedish Sign Language Corpus is also an important resource, but it is of a very different kind than the dictionary. Compiled during the years 2009–2011, the corpus consists of video recorded conversations among 42 informants aged between 20 and 82, from three separate regions in Sweden. With 14 % of the corpus having been annotated with glosses for signs, it comprises total of approximately 3,600 different signs occurring about 25,500 times (tokens) in the 42 annotated sign language discourses/video files. As these two resources sprang from different starting points, they are independent from each other; however, in the late phases of building the corpus the importance of combining work from the two became evident. This presentation will show the development of these two resources and the advantages of combining them.

English-ASL Gloss Parallel Corpus 2012: ASLG-PC12

Achraf Othman and Mohamed Jemni

A serious problem facing the Community for researchers in the field of sign language is the absence of a large parallel corpus for signs language. The ASLG-PC12 project proposes a rule-based approach for building big parallel corpus between English written texts and American Sign Language Gloss. We present a novel algorithm which transforms an English part-of-speech sentence to ASL gloss. This project was started in the beginning of 2011, a part of the project WebSign, and it offers today a corpus containing more than one hundred million pairs of sentences between English and ASL gloss. It is available online for free in order to develop and design new algorithms and theories for American Sign Language processing, for example statistical machine translation and any related fields. In this paper, we present tasks for generating ASL sentences from the corpus Gutenberg Project that contains only English written texts.

Compiling the Slovene Sign Language Corpus

Špela Vintar, Boštjan Jerko and Marjetka Kulovec

We report on the project of compiling the first corpus of the Slovene Sign Language. The paper describes the procedures of data collection, the decisions regarding informant selection and plans for transcription and annotation. We outline the particularities of the Slovene situation, especially the high variability of the language, issues concerning language competence and the attitutes of the deaf community towards such data collection. At the time of writing, the data collection stage is nearly finished with over 70 persons recorded, and trancriptions with iLex are underway. The aim of the project is to use the corpus for explorations into the grammatical properties of SSL.

Session D: Issues in the Construction of Sign Corpora

Sunday 27 May, 16:30 – 19:00
Chairperson: Stavroula-Evita Fotinea

Challenges in Development of the American Sign Language Lexicon Video Dataset (ASLLVD) Corpus

Carol Neidle, Ashwin Thangali and Stan Sclaroff

The American Sign Language Lexicon Video Dataset (ASLLVD) consists of videos of >3,300 ASL signs in citation form, each produced by 1-6 native ASL signers, for a total of almost 9,800 tokens. This dataset, including multiple synchronized videos showing the signing from different angles, will be shared publicly once the linguistic annotations and verifications are complete. Linguistic annotations include gloss labels, sign start and end time codes, start and end handshape labels for both hands, morphological and articulatory classifications of sign type. For compound signs, the dataset includes annotations for each morpheme. To facilitate computer vision-based sign language recognition, the dataset also includes numeric ID labels for sign variants, video sequences in uncompressed-raw format, camera calibration sequences, and software for skin region extraction. We discuss here some of the challenges involved in the linguistic annotations and categorizations. We also report an example computer vision application that leverages the ASLLVD: the formulation employs a HandShapes Bayesian Network (HSBN), which models the transition probabilities between start and end handshapes in monomorphemic lexical signs. Further details and statistics for the ASLLVD dataset, as well as information about annotation conventions, are available from http://www.bu.edu/asllrp/lexicon.

SignWiki – An Experiment in Creating a User-based Corpus

Sonja Erlenkamp and Olle Eriksen

In comparison to other signed languages, Norwegian Sign Language (NTS) is not well researched and documented while at the same time the need for documentation of NTS in a corpus based dictionary has been apparent to the field for quite a while. Despite some high quality applications to raise funding for corpus work, the field in Norway has not succeeded to gain enough understanding in governmental research funding institutions for the need of a corpus based dictionary, mainly because of the rather small population of NTS users. As a result, a new approach is used by involving the NTS community to create a database of signs, including their use, distribution and as far as possible other metadata. Tegnwiki (=Signwiki) is a first attempt at creating a user-based database of NTS by allowing users to contribute videos and information on isolated signs on a Wikiplatform. Like Wikipedia, the SignWiki will be open accessible, but administered by a group of experts. Obviously a SignWiki cannot replace a scientific corpus. But if this experiment is successful it might be a good starting point for countries with no or little funding for corpus projects where involvement of users is the key factor.

Where Does a Sign Start and End? Segmentation of Continuous Signing

Thomas Hanke, Silke Matthes, Anja Regen and Satu Worseck

There are two basic approaches how to segment continuous signing into individual signs:

- A sign starts where the preceding one ends (i.e. fluent signing means there are no gaps between signs)
- Transitional movements between signs do not count as part of either sign. Therefore, usually there are gaps between two signs during which the articulators move from the end of one sign to the beginning of the next.

Both approaches have their pros and cons. However, in the context of the DGS Corpus and the Dicta-Sign project the second approach offers advantages for the subsequent processing. Here we investigate how sensitive this approach is with respect to higher video frame rates.

Transcribing and Evaluating Language Skills of Deaf Children in a Multimodal and Bilingual Way: the Sensitive Issue of the Gesture/Signs Dynamics

Isabelle Estève

Transcribing and evaluating the narrative productions of 6 to 12 year-olds deaf children in their multimodal and bilingual dimensions confront us to the central question of gestures/signs distinction. This paper aims to discuss how the narrative skills of 30 deaf children schooled in different education settings – oralist, bilingual and “mixed” – led us to create transcription/annotation tools in ELAN allowing to take into account the dynamics between verbal and non-verbal material involving especially within the gestural modality. We will focus on two central points of our reflections. How to delimit productions in units into taking into account the semiotic and the structural dynamics aspects of production? How to describe and categorize the gestural processes non systematized in a linguistic form to report the developmental dynamics?

Sign Language Documentation in the Asia-Pacific Region: A Deaf-centred approach

Felix Sze, James Woodward, Gladys Tang, Jafi Lee, Ka-Yiu Cheng and Joe Ma

In this paper, we would like to share our experience in training up Deaf individuals from the Asian-Pacific countries to compile sign language dictionaries and conduct sign language research through the ‘Asia-Pacific Sign Linguistics Research and Training Program’ at the Chinese University of Hong Kong. The program, fully funded by the Nippon Foundation, is a multi-country, multi-phase project which aims at nurturing Deaf people to become sign language researchers through a series of credit-bearing training programs at the diploma and higher diploma levels. The training covers three major areas: Sign Linguistics, Sign Language Teaching and English Literacy. One important part of the training involves the production of sample dictionaries of the Deaf trainees’ own sign languages. To confirm the dictionary entries, the Deaf trainees conduct surveys in the Deaf communities in their home countries from time to time and as a result a substantial amount of lexical variants have been collected. An online database, called the Asian SignBank, is now being developed to house these lexical data and facilitate further research. Apart from basic search functions, the SignBank also incorporates detailed phonetic features of individual signs and a materials-generating function which allows quicker production of dictionaries in the future.

Natural Language Processing for Improving Textual Accessibility (NLP4ITA)

Sunday, May 27, 2012

ABSTRACTS

Editors:

Luz Rello, Horacio Saggion

Workshop Programme

9:00 – 9:10 Introduction by Workshop Chair

9:10 – 10:10 **Invited Talk** by Ruslan Mitkov
NLP and Language Disabilities

Session: Simplification

10:10 - 10:30 María Jesús Aranzabe, Arantza Díaz de Ilarraza, Itziar Gonzalez-Dios
First Approach to Automatic Text Simplification in Basque

10:30 - 11:00 Coffee break

11:00 - 11:20 Alejandro Mosquera, Elena Lloret, Paloma Moreda
Towards Facilitating the Accessibility of Web 2.0 Texts through Text Normalization Resources

Session: Resources

11:20 - 11:40 Sanja Štajner, Richard Evans, Constantin Orasan, and Ruslan Mitkov
What can readability measures really tell us about text complexity?
12:00 - 12:20 Luz Rello, Ricardo Baeza-Yates, Horacio Saggion, Jennifer Pedler
A First Approach to the Creation of a Spanish Corpus of Dyslexic Texts

Session: Vocal Aid

12:20 - 12:40 Janneke van de Loo, Guy De Pauw, Jort F. Gemmeke, Peter Karsmakers, Bert Van Den Broeck, Walter Daelemans, Hugo Van hamme
Towards Shallow Grammar Induction for an Adaptive Assistive Vocal Interface: a Concept Tagging Approach

12:40 - 13:00 Tiberiu Boroș, Dan Ștefănescu, Radu Ion
Bermuda, a data-driven tool for phonetic transcription of words

13:00 End of the Workshop

Workshop Organizers

Ricardo Baeza-Yates	Universitat Pompeu Fabra, Yahoo!
Paloma Moreda	Universidad de Alicante
Luz Rello	Universitat Pompeu Fabra
Horacio Saggion	Universitat Pompeu Fabra
Lucia Specia	University of Sheffield

Workshop Programme Committee

Sandra Aluisio	University of Sao Paulo
Ricardo Baeza-Yates	Universitat Pompeu Fabra, Yahoo!
Delphine Bernhard	University of Strassbourg
Nadjet Bouayad-Agha	Universitat Pompeu Fabra
Richard Evans	University of Wolverhampton
Caroline Gasperin	TouchType Ltd
Pablo Gervás	Universidad Complutense de Madrid
José Manuel Gómez	Universidad de Alicante
Simon Harper	University of Manchester
David Kauchak	Middlebury College
Guy Lapalme	University of Montreal
Elena Lloret	Universidad de Alicante
Paloma Martínez	Universidad Carlos III de Madrid
Aurelien Max	Paris 11
Kathleen F. McCoy	University of Delaware
Ornella Mich	Foundazione Bruno Kessler
Ruslan Mitkov	University of Wolverhampton
Paloma Moreda	Universidad de Alicante
Constantin Orasan	University of Wolverhampton
Luz Rello	Universitat Pompeu Fabra
Horacio Saggion	Universitat Pompeu Fabra
Advaith Siddharthan	University of Aberdeen
Lucia Specia	University of Sheffield
Juan Manuel Torres Moreno	University of Avignon
Markel Vigo	University of Manchester
Leo Wanner	Universitat Pompeu Fabra
Yeliz Yesilada	Middle East Technical University Northern Cyprus Campus

Preface

In recent years there has been an increasing interest in accessibility and usability issues. This interest is mainly due to the greater importance of the Web and the need to provide equal access and equal opportunity to people with diverse disabilities. The role of assistive technologies based on language processing has gained importance as it can be observed from the growing number of efforts (United Nations declarations on universal access to information or WAI guidelines related to content) and research in conferences and workshops (W4A, ICCHP, ASSETS, SLPAT, etc.). However, language resources and tools to develop assistive technologies are still scarce.

This workshop Natural Language Processing for Improving Textual Accessibility (NLP4ITA) aimed to bring together researchers focused on tools and resources for making textual information more accessible to people with special needs including diverse ranges of hearing and sight disabilities, cognitive disabilities, elderly people, low-literacy readers and adults being alphabetized, among others.

NLP4ITA had and acceptance rate of 54%, we received 11 papers from which 6 papers were accepted. We believe the accepted paper 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 Ruslan Mitkov 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 Sandra Szasz for designing and updating NLP4ITA website and to the LREC organizers. Last but not least we would like to thank our authors and the participants of the workshop.

Luz Rello and Horacio Saggion
Barcelona, 2012

Simplification

Sunday, May 27, 2012, 10:10 – 11:20
Chairperson: Luz Rello

First Approach to Automatic Text Simplification in Basque

María Jesús Aranzabe, Arantza Díaz de Ilarraza, Itziar Gonzalez-Dios

Analysis of long sentences are source of problems in advanced applications such as machine translation. With the aim of solving these problems in advanced applications, we have analysed long sentences of two corpora written in Standard Basque in order to make syntactic simplification. The result of this analysis leads us to design a proposal to produce shorter sentences out of long ones. In order to perform this task we present an architecture for a text simplification system based on previously developed general coverage tools (giving them a new utility) and on hand written rules specific for syntactic simplification. Being Basque an agglutinative language this rules are based on morphological features. In this work we focused on specific phenomena like appositions, finite relative clauses and finite temporal clauses. The simplification proposed does not exclude any target audience, and the simplification could be used for both humans and machines. This is the first proposal for Automatic Text simplification and opens a research line for the Basque language in NLP.

Towards Facilitating the Accessibility of Web 2.0 Texts through Text Normalization Resources

Alejandro Mosquera, Elena Lloret, Paloma Moreda

The Web 2.0, through its different platforms, such as blogs, social networks, microblogs, or forums allows users to freely write content on the Internet, with the purpose to provide, share and use information. However, the non-standard features of the language used in Web 2.0 publications can make social media content less accessible than traditional texts. For this reason we propose TENOR, a multilingual lexical approach for normalising Web 2.0 texts. Given a noisy sentence either in Spanish or English, our aim is to transform it into its canonical form, so that it can be easily understood by any person or text simplification tools. Our experimental results show that TENOR is an adequate tool for this task, facilitating text simplification with current NLP tools when required and also making Web 2.0 texts more accessible to people unfamiliar with these text types.

Resources

Sunday, May 27, 2012, 11:20 – 12:20
Chairperson: Horacio Saggion

What can readability measures really tell us about text complexity?

Sanja Štajner, Richard Evans, Constantin Orasan, Ruslan Mitkov

This study presents the results of an initial phase of a project seeking to convert texts into a more accessible form for people with autism spectrum disorders by means of text simplification technologies. Random samples of Simple Wikipedia articles are compared with texts from News, Health, and Fiction genres using four standard readability indices (Kincaid, Flesch, Fog and SMOG) and sixteen linguistically motivated features. The comparison of readability indices across the four genres indicated that the Fiction genre was relatively easy whereas the News genre was relatively difficult to read. The correlation of four readability indices was measured, revealing that they are almost perfectly linearly correlated and that this correlation is not genre dependent. The correlation of the sixteen linguistic features to the readability indices was also measured. The results of these experiments indicate that some of the linguistic features are well correlated with the readability measures and that these correlations are genre dependent. The maximum correlation was observed for fiction.

A First Approach to the Creation of a Spanish Corpus of Dyslexic Texts

Luz Rello, Ricardo Baeza-Yates, Horacio Saggion, Jennifer Pedler

Corpora of dyslexic texts are valuable for studying dyslexia and addressing accessibility practices, among others. However, due to the difficulty of finding texts written by dyslexics, these kinds of resources are scarce. In this paper, we introduce a small Spanish corpus of dyslexic texts with annotated errors. Since these errors require non-standard annotation, we present the annotation criteria established for the different types of dyslexic errors. We compare our preliminary findings with a similar corpus in English. This comparison suggests that the corpus shall be enlarged in future work.

Vocal Aid

Sunday, May 27, 2012, 12:20 – 13:00
Chairperson: Luz Rello

Towards Shallow Grammar Induction for an Adaptive Assistive Vocal Interface: a Concept Tagging Approach

Janneke van de Loo, Guy De Pauw, Jort F. Gemmeke, Peter Karsmakers, Bert Van Den Broeck, Walter Daelemans, Hugo Van hamme

This paper describes research within the ALADIN project, which aims to develop an adaptive, assistive vocal interface for people with a physical impairment. One of the components in this interface is a self-learning grammar module, which maps a user's utterance to its intended meaning. This paper describes a case study of the learnability of this task on the basis of a corpus of commands for the card game 'patience'. The collection, transcription and annotation of this corpus is outlined in this paper, followed by results of preliminary experiments using a shallow concept-tagging approach. Encouraging results are observed during learning curve experiments, that gauge the minimal amount of training data needed to trigger accurate concept tagging of previously unseen utterances.

Bermuda, a data-driven tool for phonetic transcription of words

Tiberiu Boroș, Dan Ștefănescu, Radu Ion

The article presents the Bermuda component of the NLPUF text-to-speech toolbox. Bermuda performs phonetic transcription for out-of-vocabulary words using a Maximum Entropy classifier and a custom designed algorithm named DLOPS. It offers direct transcription by using either one of the two available algorithms, or it can chain either algorithm to a second layer Maximum Entropy classifier designed to correct the first-layer transcription errors. Bermuda can be used outside of the NLPUF package by itself or to improve performance of other modular text-to-speech packages. The training steps are presented, the process of transcription is exemplified and an initial evaluation is performed. The article closes with usage examples of Bermuda.

Workshop on Creating Cross-language Resources for Disconnected Languages and Styles

27 May 2012

ABSTRACTS

Editors:

Patrik Lambert, Marta R. Costa-jussà, Rafael E. Banchs

Workshop Programme

14:10 – 14:20 Workshop Presentation

14:20 – 15:00 Session 1: Multilingual Database Generation

Anna Vacalopoulou, Voula Giouli, Eleni Efthimiou and Maria Giagkou, *Bridging the gap between disconnected languages: the eMiLang multi-lingual database*

Sadaf Abdul-Rauf, Mark Fishel, Patrik Lambert, Sandra Noubours and Rico Sennrich, *Extrinsic Evaluation of Sentence Alignment Systems*

15:00 – 16:00 Session 2: Cross-language Resource Derivation

Carolina Scarton and Sandra Aluísio, *Towards a cross-linguistic VerbNet-style lexicon to Brazilian Portuguese*

Parth Gupta, Khushboo Singhal and Paolo Rosso, *Multiword Named Entities Extraction from Cross-Language Text Re-use*

Carlos Rodríguez-Penagos, Jens Grivolla and Joan Codina-Filbá, *Projecting Opinion Mining resources across languages and genres*

16:00 – 16:30 Coffee break

16:30 – 17:30 Session 3: European Projects for Cross-language Resources

Arda Tezcan, Joeri Van de Walle and Heidi Depraetere, *Bologna Translation Service: Constructing Language Resources in the Educational Domain*

Frédérique Segond, Eduard Barbu, Igor Barsanti, Bogomil Kovachev, Nikolaos Lagos, Marco Trevisan and Ed Vald, *From scarcity to bounty: how Galateas can turn your scarce short queries into gold*

Tatiana Gornostay, Anita Gojun, Marion Weller, Ulrich Heid, Emmanuel Morin, Beatrice Daille, Helena Blancafort, Serge Sharoff and Claude Méchoulam, *Terminology Extraction, Translation Tools and Comparable Corpora: TTC concept, midterm progress and achieved results*

17:30 – 18:15 Panel Discussion

18:15 End of Workshop

Workshop Organizers

Patrik Lambert	University of Le Mans, France
Marta R. Costa-jussà	Barcelona Media Innovation Center, Spain
Rafael E. Banchs	Institute for Infocomm Research, Singapore

Workshop Programme Committee

Iñaki Alegria	University of the Basque Country, Spain
Marianna Apidianaki	LIMSI-CNRS, Orsay, France
Victoria Arranz	ELDA, Paris, France
Jordi Atserias	Yahoo! Research, Barcelona, Spain
Gareth Jones	Dublin City University, Ireland
Min-Yen Kan	National University of Singapore
Philipp Koehn	University of Edinburgh, UK
Udo Kruschwitz	University of Essex, UK
Yanjun Ma	Baidu Inc. Beijing, China
Sara Morrissey	Dublin City University, Ireland
Maja Popovic	DFKI, Berlin, Germany
Paolo Rosso	Universidad Politécnica de Valencia, Spain
Marta Recasens	Stanford University, USA
Wade Shen	Massachusetts Institute of Technology, Cambridge, USA

Introduction

Linguistic resources have become incredibly valuable as current computational power and data storage capacity have allowed for implementing data-driven and statistical approaches for almost any Natural Language Processing application. Empirical evidence has demonstrated, in a large number of cases and applications, how the availability of appropriate datasets can boost the performance of processing methods and analysis techniques. In this scenario, the availability of data is playing a fundamental role in a new generation of Natural Language Processing applications and technologies.

Nevertheless, there are specific applications and scenarios for which linguistic resources still continue to be scarce. Both, the diversity of languages and the emergence of new communication media and stylistic trends, are responsible for the scarcity of resources in the case of some specific tasks and applications.

In this sense, CREDISLAS aims at studying methods, developing strategies and sharing experiences on creating resources for reducing the linguistic gaps for those specific languages and applications exhibiting resource scarcity problems. More specifically, we focus our attention in three important problems:

- Minority Languages, for which scarcity of resources is a consequence of the minority nature of the language itself. In this case, attention is focused on the development of both monolingual and cross-lingual resources. Some examples in this category include: Basque, Pashto and Haitian Creole, just to mention a few.

- Disconnected Languages, for which a large amount of monolingual resources are available, but due to cultural, historical and/or geographical reasons cross-language resources are actually scarce. Some examples in this category include language pairs such as Chinese and Spanish, Russian and Portuguese, and Arabic and Japanese, just to mention a few.

- New Language Styles, which represent different communication forms or emerging stylistic trends in languages for which the available resources are practically useless. This case includes the typical examples of tweets and chat speak communications, as well as other informal form of communications, which have been recently propelled by the growing phenomenon of the Web2.0.

We hope the CREDISLAS initiative to nourish future research as well as resource development for several useful Natural Language Processing applications and technologies, which should contribute towards a richer heritage of language diversity and availability of linguistics resources for the Natural Language Processing scientific community.

With best regards,

The CREDISLAS organizing team

Patrik Lambert, University of Le Mans
Marta R. Costa-jussà, Barcelona Media Innovation Centre
Rafael E. Banchs, Institute for Infocomm Research

Session 1: Multilingual Database Generation

14:20 – 15:00
Chair: Carlos Rodríguez-Penagos

Bridging the gap between disconnected languages: the eMiLang multi-lingual database

Anna Vacalopoulou, Voula Giouli, Eleni Efthimiou and Maria Giagkou

We present a multi-lingual Lexical Resource (LR) developed in the context of a lexicographic project that involves the development of user-oriented dictionaries for immigrants in Greece. The LR caters to languages that as of yet remain disconnected, and also encompasses a variety of styles that are relevant to communicative situations that the target group is most likely to cope with. We are currently in the process of testing the feasibility to exploit this cross-language and cross-style LR for the automatic acquisition of further large-scale LRs (i.e., comparable corpora), the ultimate goal being to reduce the linguistic gap between the specific disconnected languages and styles.

Extrinsic Evaluation of Sentence Alignment Systems

Sadaf Abdul-Rauf, Mark Fishel, Patrik Lambert, Sandra Noubours and Rico Sennrich

Parallel corpora are usually a collection of documents which are translations of each other. To be useful in NLP applications such as word alignment or machine translation, they first have to be aligned at the sentence level. This paper is a user study briefly reviewing several sentence aligners and evaluating them based on the performance achieved by the SMT systems trained on their output. We conducted experiments on two language pairs and showed that using a more advanced sentence alignment algorithm may yield gains of 0.5 to 1 BLEU points.

Session 2: Cross-language Resource Derivation

15:00 – 16:00
Chair: Patrik Lambert

Towards a cross-linguistic VerbNet-style lexicon to Brazilian Portuguese

Carolina Scarton and Sandra Aluísio

This paper presents preliminary results of the Brazilian Portuguese Verbnet (VerbNet.Br). This resource is being built by using other existing Computational Lexical Resources via a semi-automatic method. We identified, automatically, 5688 verbs as candidate members of VerbNet.Br, which are distributed in 257 classes inherited from VerbNet. These preliminary results give us some directions of future work and, since the results were automatically generated, a manual revision of the complete resource is highly desirable.

Multiword Named Entities Extraction from Cross-Language Text Re-use

Parth Gupta, Khushboo Singhal and Paolo Rosso

In practice, many named entities (NEs) are multiword. Most of the research, done on mining the NEs from the comparable corpora, is focused on the single word transliterated NEs. This work presents an approach to mine Multiword Named Entities (MWNEs) from the text re-use document pairs. Text re-use, at document level, can be seen as noisy parallel or comparable text based on the level of obfuscation. Results, reported for Hindi-English language pair, are very encouraging. The approach can easily be extended to any language pair.

Projecting Opinion Mining resources across languages and genres

Carlos Rodríguez-Penagos, Jens Grivolla and Joan Codina-Filbá

Creating language-specific resources to mine opinions in user-generated content can be a laborious task, but even less funded languages have the need for such processing in our increasingly connected world. We describe some experiments in creating Catalan polar lexicons from Spanish resources using automatic word-by-word translation as well as whole corpus Machine Translation for applying bayesian classification methods. Even though some challenges remain in data sparseness and domain adaptation, we believe a practical way of transporting attitude-related contextual information is possible, beyond the more conventional translation of literal lexical meaning.

Session 3: European Projects for Cross-language Resources

16:30 – 17:30
Chair: Paolo Rosso

Bologna Translation Service: Constructing Language Resources in the Educational Domain

Arda Tezcan, Joeri Van de Walle and Heidi Depraetere

BTS – Bologna Translation Service – is an ICT PSP EU-funded project which specialises in the automatic translation of study programmes. At the core of the BTS framework are several machine translation (MT) engines through which web-based translation services are offered. Statistical machine translation (SMT) systems form the backbones for all BTS language pairs and for such systems the importance of monolingual and bilingual corpora is undeniable. Unfortunately the lack of readily available domain-specific linguistic resources for various language pairs is one of the major obstacles to build engines with high quality output. In this paper, we report on the ongoing work of language resource construction in the educational domain, focusing on various aspects of this work within the scope of BTS. We also present other relevant BTS components and methods that are used with the aim of exploiting the collected resources and improving MT quality further in the BTS framework.

From scarcity to bounty: how Galateas can turn your scarce short queries into gold

Frédérique Segond, Eduard Barbu, Igor Barsanti, Bogomil Kovachev, Nikolaos Lagos, Marco Trevisan and Ed Vald

With the growth of digital libraries and the digital library federation in addition to partially unstructured collections of documents such as web sites, a large set of vendors are offering engines for retrieving content and metadata via search requests by the end user (queries). In most cases these queries are short unstructured fragments of text in different languages that are difficult to make sense of because of the lack of context. When attempting to perform automatic translation of these queries, using machine learning approaches, the problem becomes worse as aligned corpora are almost inexistent for such types of linguistic data. The GALATEAS European project concentrates on analyzing language-based information from transaction logs and facilitates the development of improved navigation and search technologies for multilingual content access, in order to offer digital content providers an innovative approach to understanding users' behaviour.

Terminology Extraction, Translation Tools and Comparable Corpora: TTC concept, midterm progress and achieved results

Tatiana Gornostay, Anita Gojun, Marion Weller, Ulrich Heid, Emmanuel Morin, Beatrice Daille, Helena Blancafort, Serge Sharoff and Claude Méchoulam

The TTC project (Terminology Extraction, Translation Tools and Comparable Corpora) has contributed to leveraging computer-assisted translation tools, machine translation systems and multilingual content (corpora and terminology) management tools by generating bilingual terminologies automatically from comparable corpora in seven EU languages, as well as Russian and Chinese. This paper presents the main concept of TTC, discusses the issue of parallel corpora scarceness and potential of comparable corpora, and briefly describes the TTC terminology extraction workflow. The TTC terminology extraction workflow includes the collection of domain-specific comparable corpora from the web, extraction of monolingual terminology in the two domains of wind energy and mobile technology, and bilingual alignment of extracted terminology. We also present TTC usage scenarios, the way in which the project deals with under-resourced and disconnected languages, and report on the project midterm progress and results achieved during the two years of the project. And finally, we touch upon the problem of under-resourced languages (for example, Latvian) and disconnected languages (for example, Latvian and Russian) covered by the project.

Semantic Processing of Legal Texts (SPLeT-2012)

27 May 2012

ABSTRACTS

Editors:

Enrico Francesconi, Simonetta Montemagni, Wim Peters, Adam Wyner

Workshop Programme

9:00-11:30 – General Session

9:00 –09:10 – Introduction by Workshop Chairs

9:10 –09:40
Giulia Venturi, *Design and Development of TEMIS: a Syntactically and Semantically Annotated Corpus of Italian Legislative Texts*

9:40 –10:10
Guido Boella, Luigi Di Caro, Llio Humphreys, Livio Robaldo, *Using Legal Ontology to Improve Classification in the Eunomos Legal Document and Knowledge Management System*

10:10–10:30
Antonio Lazari, Mª Ángeles Zarco-Tejada, *JurWordNet and FrameNet Approaches to Meaning Representation: a Legal Case Study*

10:30 – 11:00 Coffee break

11:00–11:30
Lorenzo Bacci, Enrico Francesconi, Maria Teresa Sagri, *A Rule-based Parsing Approach for Detecting Case Law References in Italian Court Decisions*

11:30-12:30 – Session on the Results of the "Collaborative Annotation Exercise"

11:30–12:30
Adam Wyner, Wim Peters, *Semantic Annotations for Legal Text Processing using GATE Teamware*

12:30-13:00 – Position Papers Session

12:30–12:45
Paulo Quaresma, *Legal Information Extraction ← Machine Learning Algorithms + Linguistic Information*

12:45–13:00
Adam Wyner, *Problems and Prospects in the Automatic Semantic Analysis of Legal Texts*

13:00 – 14:00 Lunch break

14:00-15:45 – Session on the Results of the "First Shared Task on Dependency Parsing of Legal Texts"

14:00 –14:20
Felice Dell'Orletta, Simone Marchi, Simonetta Montemagni, Barbara Plank, Giulia Venturi, *The SPLeT--2012 Shared Task on Dependency Parsing of Legal Texts*

14:20 –14:50
Giuseppe Attardi, Daniele Sartiano and Maria Simi, *Active Learning for Domain Adaptation of Dependency Parsing on Legal Texts*

14:50 –15:10
Alessandro Mazzei, Cristina Bosco, *Simple Parser Combination*

15:10–15:30
Niklas Nisbeth, Anders Søgaard, *Parser combination under sample bias*

15:30 –15:45 – Discussion of Shared Task Results

15:45 –16:00 – Closing Remarks

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
Adam Wyner	Department of Computer Science, University of Liverpool, UK

Workshop Programme Committee

Kevin Ashley	University of Pittsburgh, USA
Johan Bos	University of Groningen, The Netherlands
Danièle Bourcier	Humboldt Universität, Berlin, Germany
Jack Conrad	Thomson-Reuters, USA
Matthias Grabmair	University of Pittsburgh, USA
Antonio Lazari	Scuola Superiore S.Anna, Pisa, Italy
Alessandro Lenci	Dipartimento di Linguistica, Università di Pisa, Italy
Leonardo Lesmo	Dipartimento di Informatica, Università di Torino, Italy
Thorne McCarty	Reutgers University, USA
Raquel Mochales Palau	Nuance International, Belgium
Paulo Quaresma	Universidade de Évora, Portugal
Tony Russell-Rose	UXLabs, UK
Erich Schweighofer	Universität Wien, Rechtswissenschaftliche Fakultät, Wien, Austria
Rolf Schwitter	Macquarie University, Australia
Manfred Stede	University of Potsdam, Germany
Daniela Tiscornia	Istituto di Teoria e Tecniche dell'Informazione Giuridica del CNR, Florence, Italy
Tom van Engers	Leibniz Center for Law, University of Amsterdam, The Netherlands
Giulia Venturi	Scuola Superiore S.Anna, Pisa, Italy
Vern R. Walker	Hofstra University School of Law, Hofstra University, USA
Stephan Walter	Germany
Radboud Winkels	Leibniz Center for Law, University of Amsterdam, The Netherlands

SPLeT-2012 Shared Task Organizers

Felice Dell'Orletta	Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR, Pisa, Italy
Simone Marchi	Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR, Pisa, Italy
Simonetta Montemagni	Istituto di Linguistica Computazionale "Antonio Zampolli" del CNR, Pisa, Italy
Barbara Plank	DISI, Università di Trento, Italy
Giulia Venturi	Scuola Superiore S.Anna, Pisa, Italy

Preface

The legal domain represents a primary candidate for web-based information distribution, exchange and management, as testified by the numerous e-government, e-justice and e-democracy initiatives worldwide. The last few years have seen a growing body of research and practice in the field of Artificial Intelligence and Law which addresses a range of topics: automated legal reasoning and argumentation, semantic and cross-language legal information retrieval, document classification, legal drafting, legal knowledge discovery and extraction, as well as the construction of legal ontologies and their application to the law domain. In this context, it is of paramount importance to use Natural Language Processing techniques and tools that automate and facilitate the process of knowledge extraction from legal texts.

Since 2008, the SPLeT 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. Within the Artificial Intelligence and Law community, there have also been a number of dedicated workshops and tutorials specifically focussing on different aspects of semantic processing of legal texts at conferences such as JURIX-2008, ICAIL-2009, ICAIL-2011, as well as in the International Summer School "Managing Legal Resources in the Semantic Web" (2007, 2008, 2009, 2010, 2011).

To continue this momentum and to advance research, a 4th Workshop on "Semantic Processing of Legal Texts" was organized at the LREC-2012 conference to bring to the attention of the broader Language Resources/Human Language Technology 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 outcome of these interactions advance research and applications and foster interdisciplinary collaboration within the legal domain.

New to this edition of the workshop were two sub-events which were meant to provide common and consistent task definitions, datasets, and evaluation for legal-IE systems along with a forum for the presentation of varying but focused efforts on their development.

The first sub-event was a shared task specifically focusing on dependency parsing of legal texts: although this is not a domain-specific task, it is a task which creates the prerequisites for advanced IE applications operating on legal texts, which can benefit from reliable pre-processing tools. For this year our aim was to create the prerequisites for more advanced domain-specific tasks (e.g. event extraction) to be hopefully organized in future SPLeT editions. The languages dealt with have been Italian and English.

The second sub-event was an online, manual, collaborative, semantic annotation exercise, the results of which are presented and discussed at the workshop. The goals of the exercise were: (1) to gain insight on and work towards the creation of a gold standard corpus of legal documents in a cohesive domain; and (2) to test the feasibility of the exercise and to get feedback on its annotation structure and workflow. For this exercise, the language was English.

The workshop and sub-events provided an overview of the state-of-the-art in legal knowledge extraction and management, presented new research and development directions and emerging trends, and in general furthered the exchange of information regarding legal language resources and human language technologies and their applications.

The papers from the workshop and sub-events are contained in these proceedings.

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. We also would like to thank the LREC 2012 Organising Committee that made this workshop possible.

The Workshop Chairs

Enrico Francesconi
Simonetta Montemagni
Wim Peters
Adam Wyner

General Session

Sunday 27 May, 9:00 – 11:30

Design and Development of TEMIS: a Syntactically and Semantically Annotated Corpus of Italian Legislative Texts

Giulia Venturi

Abstract

Methodological issues concerning the design and the development of TEMIS, a syntactically and semantically annotated corpus of Italian legislative texts, are presented and discussed in the paper. TEMIS is a heterogeneous collection of legislative texts exemplifying different sub-varieties of Italian legal language, i.e. European, national and local texts. The whole corpus has been syntactically dependency annotated and a subset has been enriched with frame-based information by customizing the formalism of the FrameNet project. In both cases, a number of domain-specific extensions of the annotation criteria developed for the general language has been foreseen. The interest in building such a corpus stems from the increasing need for annotated collections of domain-specific texts recognized by both the Artificial Intelligence and Law (AI&Law) community and the Natural Language Processing (NLP) one. In two research communities the benefits of having a resource where both domain-specific content and its underlying linguistic structure are made explicit and aligned are widely acknowledged. To the author knowledge, this is the first annotated corpus of legal texts overtly devoted to be used for legal text processing applications based on NLP tools.

Using legal ontology to improve classification in the Eunomos legal document and knowledge management system

Guido Boella, Luigi Di Caro, Llio Humphreys, Livio Robaldo

Abstract

We focus on the classification of descriptions of legal obligations in the Legal Taxonomy Syllabus. We compare the results of classification using increasing levels of semantic information. Firstly, we use the text of the concept description, analysed via the TULE syntactic parser, to disambiguate syntactically and select informative nouns. Secondly, we add as additional features for the classifier the concepts (via their ontological ID) which have been semi-automatically linked to the text by knowledge engineers in order to disambiguate the meaning of relevant phrases which are associated to concepts in the ontology. Thirdly, we consider concepts related to the prescriptions by relations such as deontological clause and sanction.

JurWordNet and FrameNet Approaches to Meaning Representation: a Legal Case Study

Antonio Lazari, Mª Ángeles Zarco-Tejada

Abstract

This paper describes JurWordNet, FrameNet and LOIS approaches towards meaning representation regarding the concept ‘State Liability’ from a cross-linguistic and comparative perspective. Our starting point has been the lexical and conceptual mismatching of legal terms that the process of harmonization in the European Union has manifested. Our study analyzes such concept in Italian, Spanish, French and English and shows how a deeper sub-language based representation of meaning is needed to account for such phenomena. We examine the most important computational-lexical models in an attempt to identify the most suitable and appropriate approach towards lexical-conceptual mismatching of the concept ‘State liability’ in the European legal tradition. Our proposal shows a formalization of the concept in the four systems mentioned and uses semantic features to represent lexical mismatching and cultural differences. With this study we show in a systematic way the differences in legal tradition and the reasons for divergence in the judicial use of related concepts.

A Rule-based Parsing Approach for Detecting Case Law References in Italian Court Decisions

Lorenzo Bacci, Enrico Francesconi, Maria Teresa Sagri

Abstract

In this paper a procedure able to detect legal references in Italian court decisions, providing automatic document hyperlinking is described. It is based on the adoption of a naming convention for case law documents, based on the metadata typically used in citations. The parsing strategy in particular is based on regular expressions, able to extract, from legal citations, the metadata used in the adopted naming convention. In particular the parser is able to implement both the ECLI and the LEX naming conventions for case law material.

Session on the Results of the “Collaborative Annotation Exercise”

Sunday 27 May, 11:30 – 12:30

Semantic Annotations for Legal Text Processing using GATE Teamware

Adam Wyner, Wim Peters

Abstract

Large corpora of legal texts are increasing available in the public domain. To make them amenable for automated text processing, various sorts of annotations must be added. We consider semantic annotations bearing on the content of the texts - legal rules, case factors, and case decision elements. Adding annotations and developing gold standard corpora (to verify rule-based or machine learning algorithms) is costly in terms of time, expertise, and cost. To make the processes efficient, we propose several instances of GATE’s Teamware to support annotation tasks for legal rules, case factors, and case decision elements. We engage annotation volunteers (law school students and legal professionals). The reports on the tasks are to be presented at the workshop.

Position Papers Session

Sunday 27 May, 12:30 – 13:00

Legal Information Extraction ← Machine Learning Algorithms + Linguistic Information

Paulo Quaresma

Abstract

In order to automatically extract information from legal texts we propose the use of a mixed approach, using linguistic information and machine learning techniques. In the proposed architecture, lexical, syntactical, and semantical information is used as input for specialized machine learning algorithms, such as, support vector machines. This approach was applied to collections of legal documents and the preliminary results were quite promising.

Problems and Prospects in the Automatic Semantic Analysis of Legal Texts

Adam Wyner

Abstract
Legislation and regulations are expressed in natural language. Machine-readable forms of the texts may be represented as linked documents, semantically tagged text, or translation to a logic. The paper considers the latter form, which is key to testing consistency of laws, drawing inferences, and providing explanations relative to input. To translate laws to a machine-readable logic, sentences must be parsed and semantically translated. Manual translation is time and labour intensive, usually involving narrowly scoping the rules. While automated translation systems have made significant progress, problems remain. The paper outlines systems to automatically translate legislative clauses to a semantic representation, highlighting key problems and proposing some tasks to address them.

Session on the Results of the "First Shared Task on Dependency Parsing of Legal Texts"

Sunday 27 May, 14:00 – 15:45

The SPLeT--2012 Shared Task on Dependency Parsing of Legal Texts

Felice Dell'Orletta, Simone Marchi, Simonetta Montemagni, Barbara Plank, Giulia Venturi

Abstract
The 4th Workshop on "Semantic Processing of Legal Texts" (SPLeT-2012) presents the first multilingual shared task on Dependency Parsing of Legal Texts. In this paper, we define the general task and its internal organization into sub--tasks, describe the datasets and the domain--specific linguistic peculiarities characterizing them. We finally report the results achieved by the participating systems, describe the underlying approaches and provide a first analysis of the final test results.

Active Learning for Domain Adaptation of Dependency Parsing on Legal Texts

Giuseppe Attardi, Daniele Sartiano and Maria Simi

Abstract
Several techniques have been explored in the literature to achieve domain adaptation in parsing. In principle fully unsupervised methods would be preferable, but the evidence so far is that none of them is effective, except for one special case of self-training used within one step of a reranking constituency parser. For the task of domain adaptation of dependency parsing to legal text, we hence chose to use a semi-supervised technique (i.e. active learning) which has consistently proved effective in other types of domain adaptation. We report on how we used active learning, i.e. selection criteria, parameters used, to perform domain adaptation in two languages: Italian and English. The results are quite positive on Italian and less on English. We discuss possible explanations for this discrepancy.

Simple Parser Combination

Alessandro Mazzei, Cristina Bosco

Abstract
This paper presents an ensemble system for dependency parsing: three parsers are separately trained and combined by means of a majority vote. The three parsers are (1) the MATE parser [http://code.google.com/p/mate-tools/], (2) the DeSR parser [http://sites.google.com/site/desrparser/], and (3) the MALT parser [http://maltparser.org/]. The MATE, that was never used before on Italian language, drastically outperforms the other parsers in the SPLeT shared task.

Parser combination under sample bias

Niklas Nisbeth, Anders Søgaard

Abstract
Combining several parsers through voting is known to improve parsing performance and robustness in supervised parsing. The intuition behind our shared task contribution to SPLeT 2012 is that voting is particularly useful when labeled data is biased, e.g. in domain adaptation.