LREC 2016
Portorož

TENTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION

Held under the Honorary Patronage of His Excellency Mr. Borut Pahor, President of the Republic of Slovenia

MAY 23 – 28, 2016

GRAND HOTEL BERNARDIN CONFERENCE CENTRE
Portorož, SLOVENIA

WORKSHOP ABSTRACTS

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

Assistant Editors: Sara Goggi, Hélène Mazo

LREC 2016, Tenth International Conference on Language Resources and Evaluation

Title: LREC 2016 Workshop Abstracts

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

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

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

ISBN 978-2-9517408-9-1
EAN 9782951740891

TABLE OF CONTENTS

Language and Ontology (LangOnto2) & Terminology and Knowledge Structures (TermiKS)

23 May 2016

ABSTRACTS

Editors:

Larisa Grčić Simeunović, Špela Vintar, Fahad Khan, Pilar León Araúz, Pamela Faber, Francesca Fontini, Artemis Parvisi, Christina Unger

Workshop Programme

Session 1
09:00 – 09:30 – Introductory Talk
Pamela Faber, *The Cultural Dimension of Knowledge Structures*

09:30 – 10:10 – Introductory Talk
John McCrae, *Putting ontologies to work in NLP: The lemon model and its applications*

10:10 – 10:30
Gregory Grefenstette, Karima Rafes, *Transforming Wikipedia into an Ontology-based Information Retrieval Search Engine for Local Experts using a Third-Party Taxonomy*

10:30 – 11:00 Coffee break

Session 2
11:00 – 11:30
Juan Carlos Gil-Berrozpe, Pamela Faber, *Refining Hyponymy in a Terminological Knowledge Base*

11:30 – 12:00
Pilar León-Araúz, Arianne Reimerink, *Evaluation of EcoLexicon Images*

12:00 – 12:30
Špela Vintar, Larisa Grčić Simeunović, *The Language-Dependence of Conceptual Structures in Karstology*

12:30 – 13:00
Takuma Asaishi, Kyo Kageura, *Growth of the Terminological Networks in Junior-high and High School Textbooks*

13:30 – 14:00 Lunch break

Session 3
14:00 – 14:20
Gabor Melli, *Semantically Annotated Concepts in KDD's 2009-2015 Abstracts*

14:20 – 14:40
Bhaskar Sinha, Somnath Chandra, *Neural Network Based Approach for Relational Classification for Ontology Development in Low Resourced Indian Language*

14:40 – 15:00
Livy Real, Valeria de Paiva, Fabricio Chalub, Alexandre Rademaker, *Gentle with the Gentilics*

15:00 – 15:30
Dante Degl'Innocenti, Dario De Nart and Carlo Tasso, *The Importance of Being Referenced: Introducing Referential Semantic Spaces*

15:30 – 16:00
Haizhou Qu, Marcelo Sardelich, Nunung Nurul Qomariyah and Dimitar Kazakov, *Integrating Time Series with Social Media Data in an Ontology for the Modelling of Extreme Financial Events*

16:00 – 16:30 Coffee break

Session 4
16:30 – 16:50
Christian Willms, Hans-Ulrich Krieger, Bernd Kiefer, ×-*Protégé An Ontology Editor for Defining Cartesian Types to Represent n-ary Relations*

16:50 – 17:10
Alexsandro Fonseca, Fatiha Sadat and François Lareau, *A Lexical Ontology to Represent Lexical Functions*

17:10 – 17:40
Ayla Rigouts Terryn, Lieve Macken and Els Lefever, *Dutch Hypernym Detection: Does Decompounding Help?*

17:40 – 18:30 Discussion and Closing

Workshop Organizers

Fahad Khan	Istituto di Linguistica Computazionale "A. Zampolli" - CNR, Italy
Špela Vintar	University of Ljubljana, Slovenia
Pilar León Araúz	University of Granada, Spain
Pamela Faber	University of Granada, Spain
Francesca Frontini	Istituto di Linguistica Computazionale "A. Zampolli" - CNR, Italy
Artemis Parvizi	Oxford University Press
Larisa Grčić Simeunović	University of Zadar, Croatia
Christina Unger	University of Bielefeld, Germany

Workshop Programme Committee

Guadalupe Aguado-de-Cea	Universidad Politécnica de Madrid, Spain
Amparo Alcina	Universitat Jaume I, Spain
Nathalie Aussenac-Gilles	IRIT, France
Caroline Barrière	CRIM, Canada
Maja Bratanić	Institute of Croatian Language and Linguistics, Croatia
Paul Buitelaar	Insight Centre for Data Analytics, Ireland
Federico Cerutti	Cardiff University
Béatrice Daille	University of Nantes, France
Aldo Gangemi	LIPN University, ISTC-CNR Rome
Eric Gaussier	University of Grenoble, France
Emiliano Giovannetti	ILC-CNR
Ulrich Heid	University of Hildesheim, Germany
Caroline Jay	University of Manchester
Kyo Kageura	University of Tokio, Japan
Hans-Ulrich Krieger	DFKI GmbH
Roman Kutlak	Oxford University Press
Marie-Claude L'Homme	OLST, Université de Montréal, Canada
Monica Monachini	ILC-CNR
Mojca Pecman	University of Paris Diderot, France
Silvia Piccini	ILC-CNR
Yuan Ren	Microsoft China
Fabio Rinaldi	Universität Zürich, Switzerland
Irena Spasic	University of Cardiff
Markel Vigo	University of Manchester

Introductory talk I

Monday 23 May, 9:00 – 9:30

The Cultural Dimension of Knowledge Structures

Pamela Faber

Frame-Based Terminology (FBT) is a cognitive approach to terminology, which directly links specialized knowledge representation to cognitive linguistics and cognitive semantics (Faber 2011, 2012, 2014). More specifically, the FBT approach applies the notion of *frame* as a schematization of a knowledge structure, which is represented at the conceptual level and held in long-term memory and which emphasizes both hierarchical and non-hierarchical conceptual relations. Frames also link elements and entities associated with general human experience or with a particular culturally embedded scene or situation. Culture is thus a key element in knowledge structures.

Cultural frames are directly connected to what has been called 'design principle' (O'Meara and Bohnemeyer 2008), 'template', 'model', 'schema' or 'frame '(Brown 2008; Burenhult 2008, Cablitz 2008, Levinson 2008). In EcoLexicon, a frame is a representation that integrates various ways of combining semantic generalizations about one category or a group of categories. In contrast, a *template* is a representational pattern for individual members of the same category. Burenhult and Levinson (2008: 144) even propose the term, *semplate*, which refers to the cultural themes or linguistic patterns that are imposed on the environment to create, coordinate, subcategorize, or contrast categories.

Although rarely explored, cultural situatedness has an impact on semantic networks, which reflect differences between terms used in closely related language cultures. Nevertheless, the addition of a cultural component to term meaning is considerably more complicated than the inclusion of terms that designate new concepts specific to other cultures. One reason for this is that certain conceptual categories are linked, for example, to the habitat of the speakers of a language and derive their meaning from the geographic and meteorological characteristics of a given geographic area or region. This paper explains the need for typology of cultural frames or profiles linked to the most prominent semantic categories. As an example, the terms for different types of local wind are analyzed and a set of meaning parameters are established that structure and enrich the cultural schemas defining meteorological concepts. These parameters highlight the cultural dimension of wind as a meteorological force.

Introductory talk II

Monday 23 May, 9:30 – 10:10

Putting ontologies to work in NLP: The *lemon* model and its applications

John McCrae

From the early development of lexicons such as WordNet it has been a goal to record rich information about the meanings of words and how they relate. In particular, there has been an ambition to provide full and formal definitions of concepts so that they can be clearly disambiguated and understood. Moreover, it is important to be able to represent the meaning of ontological concepts relative to how they are expressed in natural language and this syntax-ontology mapping is still poorly understood. To this end, we developed the *lemon* model, firstly in the Monnet project and secondly in the context of the W3C OntoLex Community Group and I will discuss how this model can express the mapping of words into ontological contexts. This model has found application is several practical areas, which I will describe in detail, but has so not yet achieved the goal of unifying ontological reasoning with natural language processing and I will describe the next steps I envision to achieve this goal.

Session 1

Monday 23 May, 10:10 – 10:30

Transforming Wikipedia into an Ontology-based Information Retrieval Search Engine for Local Experts using a Third-Party Taxonomy

Gregory Grefenstette, Karima Rafes

Wikipedia is widely used for finding general information about a wide variety of topicss. Its vocation is not to provide local information. For example, it provides plot, cast, and production information about a given movie, but not showing times in your local movie theatre. Here we describe how we can connect local information to Wikipedia, without altering its content. The case study we present involves finding local scientific experts. Using a third-party taxonomy, independent from Wikipedia's category hierarchy, we index information connected to our local experts, present in their activity reports, and we re-index Wikipedia content using the same taxonomy. The connections between Wikipedia pages and local expert reports are stored in a relational database, accessible through as public SPARQL endpoint. A Wikipedia gadget (or plugin) activated by the interested user, accesses the endpoint as each Wikipedia page is accessed. An additional tab on the Wikipedia page allows the user to open up a list of teams of local experts associated with the subject matter in the Wikipedia page. The technique, though presented here as a way to identify local experts, is generic, in that any third party taxonomy, can be used in this to connect Wikipedia to any non-Wikipedia data source.

Session 2

Monday 23 May, 11:00 – 11:30

Refining Hyponymy in a Terminological Knowledge Base

Juan Carlos Gil-Berrozpe, Pamela Faber

Hyponymy or type_of relation is the backbone of all hierarchical semantic configurations. Although recent work has focused on other relations such as meronymy and causality, hyponymy maintains its special status since it implies property inheritance. As reflected in EcoLexicon, a multilingual terminological knowledge base on the environment, conceptual relations are a key factor in the design of an internally and externally coherent concept system. Terminological knowledge bases can strengthen their coherence and dynamicity when the set of conceptual relations is wider than the typical generic-specific and part-whole relations, which entails refining both the hyponymy and meronymy relations. This paper analyzes how hyponymy is built in the EcoLexicon knowledge base and discusses the problems that can ensue when the type_of relation is too simplistically defined or systematically represented. As a solution, this paper proposes the following: (i) the correction of property inheritance; (ii) the specification of different subtypes of hyponymy; (iii) the creation of 'umbrella concepts'. This paper focuses on the first two solutions and proposes a set of parameters that can be used to decompose hyponymy.

Session 2
Monday 23 May, 11:30 – 12:00

Evaluation of EcoLexicon Images

Pilar León-Araúz, Arianne Reimerink

The multimodal knowledge base EcoLexicon includes images to enrich conceptual description and enhance knowledge acquisition. These images have been selected according to the conceptual propositions contained in the definitional templates of concepts. Although this ensures coherence and systematic selection, the images are related to each specific concept and are not annotated according to other possible conceptual propositions contained in the image. Our aim is to create a separate repository for images, annotate all knowledge contained in each one of them and then link them to all concept entries that contain one or more of these propositions in their definitional template. This would not only improve the internal coherence of EcoLexicon but it would also improve the reusability of the selected images and avoid duplicating workload. The first step in this process and the objective of the research here described is to evaluate the images already contained in EcoLexicon to see if they were adequately selected in the first place, how knowledge is conveyed through the morphological features of the image and if they can be reused for other concept entries. This analysis has provided preliminary data to further explore how concept type, conceptual relations, and propositions affect the relation between morphological features and image types chosen for visual knowledge representation.

Session 2
Monday 23 May, 12:00 – 12:30

The Language-Dependence of Conceptual Structures in Karstology

Špela Vintar, Larisa Grčić Simeunović

We explore definitions in the domain of karstology from a cross-language perspective with the aim of comparing the cognitive frames underlying defining strategies in Croatian and English. The experiment involved the semi-automatic extraction of definition candidates from our corpora, manual selection of valid examples, identification of functional units and semantic annotation with conceptual categories and relations. Our results comply with related frame-based approaches in that they clearly demonstrate the multidimensionality of concepts and the key factors affecting the choice of defining strategy, e.g. concept category, its place in the conceptual system of the domain and the communicative setting. Our approach extends related work by applying the frame-based view on a new language pair and a new domain, and by performing a more detailed semantic analysis. The most interesting finding, however, regards the crosslanguage comparison; it seems that definition frames are language- and/or culture-specific in that certain conceptual structures may exist in one language but not the other. These results imply that a cross-linguistic analysis of conceptual structures is an essential step in the construction of knowledge bases, ontologies and other domain representations.

Session 2

Monday 23 May, 12:30 – 13:00

Growth of the Terminological Networks in Junior-high and High School Textbooks

Takuma Asaishi, Kyo Kageura

In this paper, we analyze the mode of deployment of the concepts when reading through textbooks. In order to do so, we construct terminological networks guided by the discourse, with vertices representing index terms and edges representing the co-occurrence of index terms in a paragraph. We observe the growth of the terminological networks in junior-high and high school Japanese textbooks in four domains (physics, chemistry, biology and earth science). Our primary results are as follows: (1) The rate of occurrence of new terms in high school textbooks is faster than in junior-high school textbooks regardless of the domain. (2) While several connected components grow independently in junior-high school textbooks, the largest component remains dominant in high school textbooks, regardless of the domain. (3) The average number of terms that are directly connected to a term increases, and many more terms are directly connected in high school textbooks than in junior-high school textbooks in all domains except physics. In addition, terms are indirectly connected through a few terms on average and are strongly connected in partial groups throughout the text, regardless of the domain and school level. (4) The degree of centralization (i.e., few terms are connected to many terms directly or facilitate indirect connection between terms) deteriorates regardless of the domain and school level. Keywords: complex network, terminology, textbook, knowledge, network analysis

Session 3

Monday 23 May, 14:00 – 14:20

Semantically Annotated Concepts in KDD's 2009-2015 Abstracts

Gabor Melli

We introduce a linguistic resource composed of a semantically annotated corpus and a lexicalized ontology that are interlinked on mentions of concepts and entities. The corpus contains the paper abstracts from within the proceedings of ACM's SIGKDD conferences for the years 2009 through 2015. Each abstract was internally annotated to identify the concepts mentioned within the text. Then, where possible, each mention was linked to the appropriate concept node in the ontology focused on data science topics. Together they form one of the few semantic resources within a subfield of computing science. The joint dataset enables tasks such as temporal modeling of concepts over time, and the development of semantic annotation methods for documents with a large proportion of mid-level concept mentions. Finally, the resource also prepares for the transition into semantic navigation of computing science research publications. Both resources are publicly available at gabormelli.com/Projects/kdd/.

Session 3

Monday 23 May, 14:20 – 14:40

Neural Network Based Approach for Relational Classification for Ontology Development in Low Resourced Indian Language

Bhaskar Sinha, Somnath Chandra

Processing natural language for text based information especially for low resource languages is challenging task. Feature extraction and classification for domain specific terms using Natural Language Processing (NLP) techniques such as pre-processing task, processing of semantic analysis etc., provides substantial support for better evaluation techniques for the accuracy of natural language based electronic database. In this paper we are exploring Neural Network based machine learning approach for Indian languages relation extraction. Convolution Neural Network (CNN) learning method is applied to semantic relational information extracted from domain specific terms, matching with multilingual IndoWordNet database. Results of machine leaning based techniques outperform with significant increase over other classification methods such as SVM, Normalized Web Distance (NWD) and statistical methods of evaluation. The objective of using this technique along with semantic web technology is to initiate a proof of concept for ontology generation by extraction and classification of relational information from IndoWordNet. This paper also highlights domain specific challenges in developing ontology in Indian languages.

Session 3

Monday 23 May, 14:40 – 15:00

Gentle with the Gentilics

Livy Real, Valeria de Paiva, Fabricio Chalub, Alexandre Rademaker

To get from ‘Brasília is the Brazilian capital’ to ‘Brasília is the capital of Brazil’ is obvious for a human, but it requires effort from a Natural Language Processing system, which should be helped by a lexical resource to retrieve this information. Here, we investigate how to deal with this kind of lexical information related to location entities, focusing in how to encode data about demonyms and gentilics in the Portuguese OpenWordnet-PT. Keywords: lexical resources, ontology, gentilics, Wordnet

Session 3

Monday 23 May, 15:00 – 15:30

The Importance of Being Referenced: Introducing Referential Semantic Spaces

Dante Degl’Innocenti, Dario De Nart and Carlo Tasso

The Web is constantly growing and to cope with its ever-increasing expansion semantic technologies are an extremely valuable ally. The major drawback of such technologies, however, is that providing a formal model of a domain is time consuming task that requires expert knowledge and, on the other hand, extracting semantic data from text in an automatic way, although possible, is still extremely hard since it requires extensive human-annotated training corpora and non trivial document pre-processing. In this work we introduce a vector space representation of concept associations that can be built in an unsupervised way with minimal pre-processing effort and allows for associative reasoning supporting word sense disambiguation and related entity retrieval tasks.

Session 3

Monday 23 May, 15:30 – 16:00

Integrating Time Series with Social Media Data in an Ontology for the Modelling of Extreme Financial Events

Haizhou Qu, Marcelo Sardelich, Nunung Nurul Qomariyah and Dimitar Kazakov

This article describes a novel dataset aiming to provide insight on the relationship between stock market prices and news on social media, such as Twitter. While several financial companies advertise that they use Twitter data in their decision process, it has been hard to demonstrate whether online postings can genuinely affect market prices. By focussing on an extreme financial event that unfolded over several days and had dramatic and lasting consequences we have aimed to provide data for a case study that could address this question. The dataset contains the stock market price of Volkswagen, Ford and the S&P500 index for the period immediately preceding and following the discovery that Volkswagen had found a way to manipulate in its favour the results of pollution tests for their diesel engines. We also include a large number of relevant tweets from this period alongside key phrases extracted from each message with the intention of providing material for subsequent sentiment analysis. All data is represented as a ontology in order to facilitate its handling, and to allow the integration of other relevant information, such as the link between a subsidiary company and its holding or the names of senior management and their links to other companies.

Session 4

Monday 23 May, 16:30 – 16:50

×-Protégé An Ontology Editor for Defining Cartesian Types to Represent n-ary Relations

Christian Willms, Hans-Ulrich Krieger, Bernd Kiefer

Arbitrary n-ary relations ($n \geq 1$) can, in principle, be realized through binary relations obtained by a reification process which introduces new individuals to which the additional arguments are linked via “accessor” properties. Modern ontologies which employ standards such as RDF and OWL have mostly obeyed this restriction, but have struggled with it nevertheless. In (Krieger and Willms, 2015), we have laid the foundations for a theory-agnostic extension of RDFS and OWL and have implemented in the last year an extension of Protégé, called ×-Protégé, which supports the definition of Cartesian types to represent n-ary relations and relation instances. Not only do we keep the distinction between the domain and the range of an n-ary relation, but also introduce so-called extra arguments which can be seen as position-oriented unnamed annotation properties and which are accessible to entailment rules. As the direct representation of n-ary relations abolishes RDF triples, we have backed up ×-Protégé by the semantic repository and entailment engine HFC which supports tuples of arbitrary length. ×-Protégé is programmed in Java and is made available under the Mozilla Public License.

Session 4

Monday 23 May, 16:50 – 17:10

A Lexical Ontology to Represent Lexical Functions

Alexsandro Fonseca, Fatiha Sadat, François Lareau

Lexical functions are a formalism that describes the combinatorial, syntactic and semantic relations among individual lexical units in different languages. Those relations include both paradigmatic relations, i.e. vertical or “in absence”, such as synonymy, antonymy and meronymy, and syntagmatic relations, i.e. horizontal or “in presence”, such as intensification (deeply committed), confirmative (valid argument) and support verbs (give an order, subject to an interrogation). We present in this paper a new lexical ontology, called Lexical Function Ontology (LFO), as a model to represent lexical functions. The aim is for our ontology to be combined with other lexical ontologies, such as the Lexical Model for Ontologies (lemon) and the Lexical Markup Framework (LMF), and to be used for the transformation of lexical networks into the semantic web formats, enriched with the semantic information given by the lexical functions, such as the representation of syntagmatic relations (e.g. collocations) usually absent from lexical networks.

Session 4

Monday 23 May, 17:10 – 17:40

Dutch Hypernym Detection: Does Decompounding Help?

Ayla Rigouts Terryn, Lieve Macken and Els Lefever

This research presents experiments carried out to improve the precision and recall of Dutch hypernym detection. To do so, we applied a data-driven semantic relation finder that starts from a list of automatically extracted domain-specific terms from technical corpora, and generates a list of hypernym relations between these terms. As Dutch technical terms often consist of compounds written in one orthographic unit, we investigated the impact of a decompounding module on the performance of the hypernym detection system. In addition, we also improved the precision of the system by designing filters taking into account statistical and linguistic information. The experimental results show that both the precision and recall of the hypernym detection system improved, and that the decompounding module is especially effective for hypernym detection in Dutch.

Cross-Platform Text Mining and Natural Language Processing Interoperability

23 May 2016

ABSTRACTS

Editors:

Richard Eckart de Castilho, Sophia Ananiadou, Thomas Margoni, Wim Peters, Stelios Piperidis

Workshop Programme

Opening Session 1

09.00 – 09.10 – Introduction

09.10 – 10.00 – Alessandro di Bari, *Interoperability — Can a model driven approach help to overcome organizational constraints? (invited talk)*

Session 1: Lightning talks part I

10.00 – 10.30 – Petr Knoth and Nancy Pontika, *Aggregating Research Papers from Publishers' Systems to Support Text and Data Mining: Deliberate Lack of Interoperability or Not?*

Dominique Estival, *Alveo: making data accessible through a unified interface – a pipe-dream?*

Nancy Ide, Keith Suderman, James Pustejovsky, Marc Verhagen, Christopher Cieri and Eric Nyberg, *The Language Application Grid*

Mouhamadou Ba and Robert Bossy, *Interoperability of corpus processing workflow engines: the case of AlvisNLP/ML in OpenMinTeD*

Sven Hodapp, Sumit Madan, Juliane Fluck and Marc Zimmermann, *Integration of UIMA Text Mining Components into an Event-based Asyn- chronous Microservice Architecture*

Richard Eckart de Castilho, *Interoperability = f (community, division of labour)*

10.30 – 11.00 – Coffee break

Session 2: Lightning talks part II

11.00 – 11.45 – John P. McCrae, Georgeta Bordea and Paul Buitelaar, *Linked Data and Text Mining as an Enabler for Reproducible Research*

Wim Peters, *Tackling Resource Interoperability: Principles, Strategies and Models*

Lana Yeganova, Sun Kim, Grigory Balasanov, Kristin Bennett, Haibin Liu and W. John Wilbur, *The DDINCBI Corpus — Towards a Larger Resource for Drug-Drug Interactions in PubMed*

Rodrigo Agerri, Itziar Aldabe, Egoitz Laparra, German Rigau, Antske Fokkens, Paul Huijgen, Marieke van Erp, Ruben Izquierdo Bevia, Piek Vossen, Anne-Lyse Minard and Bernardo Magnini, *Multilingual Event Detection using the NewsReader pipelines*

Shashank Sharma, PYKL Srinivas and Rakesh Balabantaray, *Emotion Detection using Online Machine Learning Method and TLBO on Mixed Script*

Gil Francopoulo, Joseph Mariani and Patrick Paroubek, *Text mining for notability computation*

Thomas Margoni and Giulia Dore, *Why We Need a Text and Data Mining Exception (but it is not enough)*

Penny Labropoulou, Stelios Piperidis, Thomas Margoni, *Legal Interoperability Issues in the Framework of the OpenMinTeD Project: a Methodological Overview*

Hege van Dijke and Stelios Piperidis, *eInfrastructures: crossing boundaries, discovering common work, achieving common goals*

Session 3: Discussion rounds I

11.45 – 12.00 – Constitution of breakout groups

12.00 – 13.00 – Breakout groups – suggested topics
- Discovery of content and access to content
- Interoperability across tools, frameworks, and platforms
- Interoperability across language resources and knowledge resources
- Legal and policy issues
- Interoperability in multi-lingual and cross-lingual scenarios
- eInfrastructures

13.00 – 14.00 – Lunch break

Session 4: Discussion rounds II

14.00 – 16.00 – Breakout groups (continued)

16.00 – 16.30 – Coffee break

Session 5: Closing session

16.30 – 18.00 – Presentation of the breakout group results
Plenary discussion

18.00 – End of workshop

Organising Committee

Richard Eckart de Castilho	Technische Universität Darmstadt, Germany
Sophia Ananiadou	University of Manchester, UK
Thomas Margoni	University of Stirling, UK
Wim Peters	University of Sheffield, UK
Stelios Piperidis	ILSP/ARC, Greece

Programme Committee

Dominique Estival	Western Sydney University, Australia
Iryna Gurevych	Technische Universität Darmstadt, Germany
Jens Grivolla	Universitat Pompeu Fabra, Spain
John Philip McCrae	National University of Ireland, Galway, Ireland
Joseph Mariani	LIMSI/CNRS, France
Kalina Bontcheva	University of Sheffield, UK
Lucie Guibault	University of Amsterdam, The Netherlands
Menzo Windhouwer	Meertens Institute, The Netherlands
Nancy Ide	Vassar College, USA
Natalia Manola	ILSP/ARC, Greece
Nicolas Hernandez	University of Nantes, France
Pei Chen	Wired Informatics, USA
Peter Klügl	Averbis GmbH, Germany
Rafal Rak	UberResearch and University of Manchester, UK
Renaud Richardet	EPFL, Switzerland
Robert Bossy	INRA, France
Thilo Götz	IBM, Germany
Steven Bethard	University of Alabama at Birmingham, USA
Torsten Zesch	University of Duisburg-Essen, Germany
Yohei Murakami	Kyoto University, Japan

Preface

Recent years have witnessed an upsurge in the quantity of available digital research data, offering new insights and opportunities for improved understanding. Following advances in Natural Language Processing (NLP), Text and data mining (TDM) is emerging as an invaluable tool for harnessing the power of structured and unstructured content and data. Hidden and new knowledge can be discovered by using TDM at multiple levels and in multiple dimensions. However, text mining and NLP solutions are not easy to discover and use, nor are they easy to combine for end users.

Multiple efforts are being undertaken world-wide to create TDM and NLP platforms. These platforms are targeted at specific research communities, typically researchers in a particular location, e.g. OpenMinTeD, CLARIN (Europe), ALVEO (Australia), or LAPPS (USA). All of these platforms face similar problems in the following areas: discovery of content and analytics capabilities, integration of knowledge resources, legal and licensing aspects, data representation, and analytics workflow specification and execution.

The goal of cross-platform interoperability raises many problems. At the level of content, metadata, language resources, and text annotations, we use different data representations and vocabularies. At the level of workflows, there is no uniform process model that allows platforms to smoothly interact. The licensing status of content, resources, analytics, and of the output created by a combination of such licenses is difficult to determine and there is currently no way to reliably exchange such information between platforms. User identity management is often tightly coupled to the licensing requirements and likewise an impediment for cross-platform interoperability.

R. Eckart de Castilho, S. Ananiadou, T. Margoni, W. Peters, S. Piperidis May 2016

Abstracts

Invited talk

Monday, 23 May 2016, 9.10 – 10.00

Interoperability— Can a model driven approach help to overcome organizational constraints?

Alessandro di Bari

Building Natural Language Processing application and services is a complex task by its very nature but also for the variety of data formats and different frameworks involved. We have to deal with several standards, different pipeline management tools and knowledge representation models. Keeping everything at the pure implementation level can be overwhelming and very costly. This is even more important in the Healthcare field, where security, privacy and compliance in general play a key role.

Therefore, as the historical evolution of Software Development shows us, rising the level of abstraction of the development languages and having the proper automation in place, can be key for a smoother and faster delivery process for NLP applications.
In this session, we will walk through the current major approaches for knowledge representation in the NLP field and we will show how a model driven approach can help in improving the interoperability among different applications and services both within and outside the boundaries of the organization.

Session 1: Lightning talks part I

Monday, 23 May 2016, 10.00 – 10.30

Aggregating Research Papers from Publishers' Systems to Support Text and Data Mining: Deliberate Lack of Interoperability or Not?

Petr Knoth and Nancy Pontika

In the current technology dominated world, interoperability of systems managed by different organisations is an essential property enabling the provision of services at a global scale. In the Text and Data Mining field (TDM), interoperability of systems offering access to text corpora offers the opportunity of increasing the uptake and impact of TDM applications. The global corpus of all research papers, i.e. the collection of human knowledge so large no one can ever read in their lifetime, represents one of the most exciting opportunities for TDM. Although the Open Access movement, which has been advocating for free availability and reuse rights to TDM from research papers, has achieved some major successes on the legal front, the technical interoperability of systems offering free access to research papers continues to be a challenge. COnnecting REpositories (CORE) aggregates the world's open access full-text scientific manuscripts from repositories, journals and publisher systems. One of the main goals of CORE is to harmonise and pre-process these data to lower the barrier for TDM. In this paper, we report on the preliminary results of an interoperability survey of systems provided by journal publishers, both open access and toll access. This helps us to assess the current level of systems' interoperability and suggest ways forward.

Alveo: making data accessible through a unified interface – a pipe-dream?

Dominique Estival

This paper addresses an old issue in corpus management which is still problematic in real-life systems: to allow users to explore and access data from various sources using a single simple interface, thus creating a tension between ease of use and over-simplification. This is then mirrored in the similar difficulty encountered with a simple data upload facility. In Alveo, the Virtual Lab for Human Communication Science, the original unified interface was sufficient for most of the datasets but proved inadequate in some cases. This paper is intended to facilitate a discussion on best practice with developers who may propose different solutions and with researchers who may have other requirements for their own datasets. We describe specific challenges posed by some datasets for Alveo, issues faced by users, identify the problems with the current state of development and propose several solutions.

The Language Application Grid

Nancy Ide, Keith Suderman, James Pustejovsky, Marc Verhagen, Christopher Cieri and Eric Nyberg

We describe the LAPPS Grid and its Galaxy front-end, focusing on its ability to interoperate between a variety of NLP platforms. The LAPPS Grid project has been a leading force in the development of specifications for web service interoperability on syntactic and semantic levels. *Syntactic interoperability* among services is enabled through LIF, the LAPPS Interchange Format, which is expressed using the JSON-LD exchange format. JSON-LD is a widely accepted format that allows data represented in the international standard JSON format to interoperate at Web-scale. *Semantic interoperability* is achieved through the LAPPS Web Service Exchange Vocabulary, which has been developed by closely with interested and invested groups to develop a lightweight, web-accessible, and readily mappable hierarchy of concepts in a bottom-up, “as needed” basis.

Interoperability of corpus processing workflow engines: the case of AlvisNLP/ML in OpenMinTeD

Mouhamadou Ba and Robert Bossy

AlvisNLP/ML is a corpus processing engine developed by the Bibliome group. It has been used in several experiments and end-user applications. We describe its design principles and data and workflow models, then we discuss interoperability challenges in the context of the OpenMinTeD project. The objective of OpenMinTeD (EC/H2020) is to create an infrastructure for Text and Data Mining (TDM) of scientific and scholarly publications. In order to offer to the infrastructure users a single entry point and the widest range of tools as possible, the major European corpus processing engines will be made interoperable, including Argo, DKPro, and GATE. We show that AlvisNLP/ML can be fully integrated into the OpenMinTeD platform while maintaining its originality.

Integration of UIMA Text Mining Components into an Event-based Asynchronous Microservice Architecture

Sven Hodapp, Sumit Madan, Juliane Fluck and Marc Zimmermann

Distributed compute resources are necessary for compute-intensive information extraction tasks processing large collections of heterogeneous documents (e.g. patents). For optimal usage of such resources, the breaking down of complex workflows and document sets into independent smaller units is required. The UIMA framework facilitates implementation of modular workflows, which represents an ideal structure for parallel processing. Although UIMA AS already includes parallel processing functionality, we tested two other approaches for distributed computing. First, we integrated UIMA workflows into the grid middleware UNICORE, which allows high performance distributed computing using control structures like loops or branching. While good distribution management and performance is a key requirement, portability, flexibility, interoperability, and easy usage are also desired features. Therefore, as an alternative, we deployed UIMA applications in a microservice architecture that supports all these aspects. We show that UIMA applications are well-suited to run in a microservice architecture while using an event-based asynchronous communication method. These applications communicate through a standardized STOMP message protocol via a message broker. Within this architecture, new applications can easily be integrated, portability is simple, and interoperability also with non-UIMA components is given. Markedly, a first test shows an increase of processing performance in comparison to the UNICORE-based HPC solution.

Interoperability = f(community; division of labour)

Richard Eckart de Castilho

This paper aims to motivate the hypothesis that practical interoperability can be seen as a function of whether and how stakeholder communities duplicate or divide work in a given area or market. We focus on the area of language processing which traditionally produces many diverse tools that are not immediately interoperable. However, there is also a strong desire to combine these tools into processing pipelines and to apply these to a wide range of different corpora. The space opened between generic, inherently “empty” interoperability frameworks that offer no NLP capabilities themselves and dedicated NLP tools gave rise to a new class of NLP-related projects that focus specifically on interoperability: *component collections*. This new class of projects drives interoperability in a very pragmatic way that could well be more successful than, e.g., past efforts towards standardised formats which ultimately saw little adoption or support by software tools.

Session 2: Lightning talks part 2

Monday, 23 May 2016, 11.00 – 11.45

Linked Data and Text Mining as an Enabler for Reproducible Research

John P. McCrae, Georgeta Bordea and Paul Buitelaar

Research data is one of the most important outcomes of many research projects and a key for enabling reproducibility in the analytic data sciences. In this paper, we explain three main challenges that complicate reproducibility namely, the difficulty of identifying datasets unambiguously, the lack of open repositories for scientific data and finally the lack of tools for understanding published science. We consider the use of linked data and text mining as two tools to solve these issues and-discuss how they may ameliorate these issues.

Tackling Resource Interoperability: Principles, Strategies and Models

Wim Peters

In order to accommodate the flexible exploitation and creation of knowledge resources in text and data mining (TDM) workflows, the TDM architecture will need to enable the re-use of resources encoding linguistic/terminological/ontological knowledge, such as ontologies, thesauri, lexical databases and the output of linguistic annotation tools. For this purpose resource interoperability is required in order to enable text mining tools to uniformly handle these knowledge resources and operationalise interoperable workflows. The Open Mining Infrastructure for Text and Data (OpenMinTeD) aims at defining this interoperability by adhering to standards for modelling and knowledge representation, and by defining a mapping structure for the harmonisation of information contained in heterogeneous resources.

The DDINCBI Corpus— Towards a Larger Resource for Drug-Drug Interactions in PubMed

Lana Yeganova, Sun Kim, Grigory Balasanov, Kristin Bennett, Haibin Liu and W. John Wilbur

Manually annotated corpora are of great importance for the development of NLP systems, both as training and evaluation data. However, the shortage of annotated corpora frequently presents a key bottleneck in the process of developing reliable applications in the health and biomedical domain and demonstrates a need for creating larger annotated corpora. Utilizing and integrating existing corpora appears to be a vital, yet not trivial, avenue towards achieving the goal. Previous studies have revealed that drug-drug interaction (DDI) extraction methods when trained on DrugBank data do not perform well on PubMed articles. With the ultimate goal of improving the performance of our DDI extraction method on PubMed(®) articles, we construct a new gold standard corpus of drug-drug interactions in PubMed that we call the DDINCBI corpus. We combine it with the existing DDIExtraction 2013 PubMed corpus and demonstrate that by merging these two corpora higher performance is achieved compared to when either source is used separately. We release the DDINCBI corpus and make it publicly available for download in BioC format at: http://bioc.sourceforge.net/. In addition, we make the existing DDIExtraction 2013 corpus available in BioC format.

Multilingual Event Detection using the NewsReader pipelines

Rodrigo Agerri, Itziar Aldabe, Egoitz Laparra, German Rigau, Antske Fokkens, Paul Huijgen, Marieke van Erp, Ruben Izquierdo Bevia, Piek Vossen, Anne-Lyse Minard and Bernardo Magnini

We describe a novel modular system for cross-lingual event extraction for English, Spanish,, Dutch and Italian texts. The system consists of a ready-to-use modular set of advanced multilingual Natural Language Processing (NLP) tools. The pipeline integrates modules for basic NLP processing as well as more advanced tasks such as cross-lingual Named Entity Linking, Semantic Role Labeling and time normalization. Thus, our cross-lingual framework allows for the interoperable semantic interpretation of events, participants, locations and time, as well as the relations between them.

Emotion Detection using Online Machine Learning Method and TLBO on Mixed Script

Shashank Sharma, PYKL Srinivas and Rakesh Balabantaray

Due to rapid modernization of our societies, most people, if not all, have access to online social media and mobile communication devices. These people hail from diverse cultures and ethnicity,

and interact with each other more often on these social media sites. Moreover, due to their distinct backgrounds, they all have an influence on the common language in which they communicate. Also, many users employ a myriad of shorthand, emoticons and abbreviations in their statements to reduce their effort. This calls for a means to assist in better communications through social media.
In our work, we have researched on understanding the underlying emotions and sentiments of these interactions and communications. Our focus was on analyzing the conversations by Indians in the code-mix of English and Hindi languages and identifying the usage patterns of various words and parts of speech. We have categorized statements into 6 groups based on emotions and improved the model using TLBO technique and online learning algorithms. These features were integrated in our application to assist the mobile device users in quickly sort and prioritize their messages based on the emotions attached with the statements and provide much more immersive communications with their friends and family.

Text mining for notability computation

Gil Francopoulo, Joseph Mariani and Patrick Paroubek

In this article, we propose an automatic computation for the notability of an author based on four criteria which are: production, citation, collaboration and innovation. The algorithms and formulas are formally presented, and then applied to a given scientific community: the Natural Language Processing (NLP) group of scientific authors gathering 48,894 people. For this purpose, a large corpus of NLP articles produced from 1965 up to 2015 has been collected and labeled as NLP4NLP with 65,003 documents. This represents a large part of the existing published articles in the NLP field over the last 50 years. The two main points of the approach are first that the computation combines pure graph algorithms and NLP systems. The second point deals with the interoperability aspects both for the corpus and the tools.

Why We Need a Text and Data Mining Exception (but it is not enough)

Thomas Margoni and Giulia Dore

Text and Data Mining (TDM) has become a key instrument in the development of scientific research. Its ability to derive new informational value from existing text and data makes this analytical tool a necessary element in the current scientific environment. TDM crucial importance is particularly evident in a historical moment when the extremely high amounts of information produced (scholarly publications, databases and datasets, social networks, etc), make it unlikely, if not impossible, for humans to read them all. Nevertheless, TDM, at least in the EU, is often a copyright infringement. This situation illustrates how certain legal provisions stifle scientific development, instead of fostering it, with significant damage for EU based researchers and research institutions and for the European socio-economic competitiveness more in general. Other countries leading the scientific and technological development have already implemented legislative or judicial solution permitting TDM, also for commercial purposes. This extended abstract suggests, as it has been already advocated in literature and in policy documents, that a mandatory TDM exception, not limited to non-commercial research, is needed to bring the EU on the same level playing field as other jurisdictions, such as the US and Japan.

Legal Interoperability Issues in the Framework of the OpenMinTeD Project: a Methodological Overview

Penny Labropoulou, Stelios Piperidis, Thomas Margoni

This paper is a first analysis of the legal interoperability issues in the framework of the OpenMinTeD (OMTD) project (www.openminted.eu), which aims to create an open, service-oriented e-Infrastructure for Text and Data Mining (TDM) of scientific and scholarly content. The paper offers an overview into the methods for achieving such interoperability.

Breakout Groups

Monday, 23 May 2016, 12.00 – 16.00

The workshop is planned as an open-space event in which the workshop participants host and participate in discussions related to the topics of interest. Based on the submissions to the workshop, we see preliminary clusters forming around the following topics:

- Discovery of content and access to content;
- Interoperability across tools, frameworks, and platforms;
- Interoperability across language resources and knowledge resources;
- Legal and policy issues;
- Interoperability in multi-lingual and cross-lingual scenarios

However, as part of pre-workshop discussions, based on the number of participants, and possibly even based on feedback generated during the workshop itself, these may be adapted.

eInfrastructures: crossing boundaries, discovering common work, achieving common goals

Monday, 23 May 2016, 12.00 – 16.00

Moderators: Stelios Piperidis (Athena Research Centre), Hege van Dijke (LIBER Europe)

The quantity and variety of digital research data of all types, the increasing number of solutions for mining such data to discover new knowledge, the variety of requirements of users of such data mining tools and services pose new challenges to the numerous infrastructural initiatives. When it comes to language data and processing, we experience multiple efforts towards generic or domain specific platforms for data aggregation and delivery, tools and services discovery, workflows orchestration and execution. Common questions in these efforts consist in interoperability at different levels: metadata of language resources and annotations, data representations and vocabularies, services across platforms and frameworks, cloud computing infrastructures, access restrictions and permissions for certain operations. In this session, we bring together strategic players in order to discuss parallel efforts being undertaken and opportunities for long-term harmonisation and strategic collaboration among existing (and future) e-Infrastructures in Europe, and the world at large, with a focus on, inter alia:

- What are the necessary strategies to enable crossing the boundaries of platforms, scientific domains, languages, national legislations?
- What can we learn from similar attempts so far?
- What can the research community do to promote these goals?
- What sort of alliances and policies are necessary to support overcoming the current barriers?

Building and Using Comparable Corpora

23 May 2016

ABSTRACTS

Editors:

Reinhard Rapp, Pierre Zweigenbaum, Serge Sharoff

Workshop Programme

Monday, May 23, 2016

09.15–9.25 ***Opening Remarks***

Session 1: Invited Presentation

09.25–10.30 Ruslan Mitkov
The Name of the Game is Comparable Corpora

10.30–11.00 ***Coffee Break***

Session 2: Building Comparable Corpora

11:00–11:30 Andrey Kutuzov, Mikhail Kopotev, Tatyana Sviridenko and Lyubov Ivanova
Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints

11:30–12:00 Yong Xu and François Yvon
A 2D CRF Model for Sentence Alignment

12:00–12:30 Mehdi Mohammadi
Parallel Document Identification using Zipf's Law

12.30–14.00 ***Lunch Break***

Session 3: Invited Presentation

14.00–15.00 Gregory Grefenstette
Exploring the Richness and Limitations of Web Sources for Comparable Corpus Research

Session 4: Applications of Comparable Corpora

15.00–15.30 Zede Zhu, Xinhua Zeng, Shouguo Zheng, Xiongwei Sun, Shaoqi Wang and Shizhuang Weng
A Mutual Iterative Enhancement Model for Simultaneous Comparable Corpora and Bilingual Lexicons Construction

15:30–16:00 Ana Sabina Uban
Hard Synonymy and Applications in Automatic Detection of Synonyms and Machine Translation

16.00–16.30 ***Coffee Break***

Session 5: Discussion

16:30–17:30 Pierre Zweigenbaum, Serge Sharoff and Reinhard Rapp
Towards Preparation of the Second BUCC Shared Task: Detecting Parallel Sentences in Comparable Corpora

17.30–17.35 ***Closing***

Workshop Organizers

Reinhard Rapp	University of Mainz, Germany
Pierre Zweigenbaum	LIMSI, CNRS, Université Paris-Saclay, Orsay, France
Serge Sharoff	University of Leeds, UK

Workshop Programme Committee

Ahmet Aker	University of Sheffield, UK
Hervé Déjean	Xerox Research Centre Europe, Grenoble, France
Éric Gaussier	Université Joseph Fourier, Grenoble, France
Vishal Goyal	Punjabi University, Patiala, India
Gregory Grefenstette	INRIA, Saclay, France
Silvia Hansen-Schirra	University of Mainz, Germany
Hitoshi Isahara	Toyohashi University of Technology
Kyo Kageura	University of Tokyo, Japan
Philippe Langlais	Université de Montrèal, Canada
Shervin Malmasi	Harvard Medical School, Boston, MA, USA
Michael Mohler	Language Computer Corp., US
Emmanuel Morin	Université de Nantes, France
Dragos Stefan Munteanu	Language Weaver, Inc., US
Lene Offersgaard	University of Copenhagen, Denmark
Ted Pedersen	University of Minnesota, Duluth, US
Reinhard Rapp	University of Mainz, Germany
Serge Sharoff	University of Leeds, UK
Michel Simard	National Research Council Canada
Pierre Zweigenbaum	LIMSI-CNRS, Orsay, France

Introduction

In the language engineering and the linguistics communities, research on comparable corpora has been motivated by two main reasons. In language engineering, on the one hand, it is chiefly motivated by the need to use comparable corpora as training data for statistical Natural Language Processing applications such as statistical machine translation or cross-lingual retrieval. In linguistics, on the other hand, comparable corpora are of interest in themselves by making possible inter-linguistic discoveries and comparisons. It is generally accepted in both communities that comparable corpora are documents in one or several languages that are comparable in content and form in various degrees and dimensions. We believe that the linguistic definitions and observations related to comparable corpora can improve methods to mine such corpora for applications of statistical NLP. As such, it is of great interest to bring together builders and users of such corpora.

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

Following the eight previous editions of the workshop which took place in Africa (LREC’08 in Marrakech), America (ACL’11 in Portland), Asia (ACL-IJCNLP’09 in Singapore and ACL-IJCNLP’15 in Beijing), Europe (LREC’10 in Malta, ACL’13 in Sofia, and LREC’14 in Reykjavik) and also on the border between Asia and Europe (LREC’12 in Istanbul), the workshop this year is co-located with LREC’16 in Portorož, Slovenia.

We would like to thank all people who in one way or another helped in making this workshop once again a success. Our special thanks go to Ruslan Mitkov and Gregory Grefenstette for accepting to give invited presentations, to the members of the program committee who did an excellent job in reviewing the submitted papers under strict time constraints, and to the LREC’16 workshop chairs and organizers. Last but not least we would like to thank our authors and the participants of the workshop.

Reinhard Rapp, Pierre Zweigenbaum, Serge Sharoff May 2016

Session 1: Invited Presentation

Monday 23 May, 9:25 – 10:30

The Name of the Game is Comparable Corpora

Ruslan Mitkov

Corpora have long been the preferred resource for a number of NLP applications and language users. They offer a reliable alternative to dictionaries and lexicographical resources which may offer only limited coverage. In the case of terminology, for instance, new terms are coined on a daily basis and dictionaries or other lexical resources, however up-to-date they are, cannot keep up with the rate of emergence of new terms. As a result, terminologists (or term extraction programs) seek to analyse the use and/or identify the translation of a specific term using corpora.

Ideally, parallel data would be the best resource both for multilingual NLP applications such as Machine Translation systems and for users such as translators, interpreters or language learners. However, parallel corpora or translation memories may not be available, they may be time-consuming to develop or difficult to acquire as they may be expensive or proprietary. An alternative and more promising approach would be to benefit from comparable corpora which are easier to compile for a specific purpose or task.

Comparable corpora, whether strictly comparable by definition or 'loosely' comparable, have already been used in applications such as Machine Translation (Rapp, Sharoff and Zeigenbaum 2016) and term extraction and have been used by translators (Corpas and Seghiri 2009). The good news is that comparable corpora can facilitate almost any multilingual application and can beneficial to almost any language user. The view of the speaker is that comparable corpora are the most versatile, valuable and practical resource for multilingual NLP. The invited talk at the BUCC workshop at LREC'2016 will show that comparable corpora can offer more in terms of value and can support a wider range of applications than has been demonstrated so far in the state of the art. The talk will present completed and ongoing work conducted by the speaker and his colleagues at the Research Group in Computational Linguistics at the University of Wolverhampton in the domain of comparable corpora. The talk will start with a discussion of the notion of comparable corpora and issues related to their use and compilation, and will briefly outline work by the speaker and his colleagues on the methodology related to the extraction of comparable documents and the building of purpose-specific comparable corpora.

Next the work carried out by the author on the automatic identification of cognates and false friends using comparable data will be presented. This will be followed by the presentation of three novel approaches developed by the speaker which use comparable data but do not resort to any dictionaries or parallel corpora, together with extensive evaluations of their performance. The speaker will then focus on the use of purpose-built comparable corpora and NLP methodology in a project whose objective was to test the validity of so-called translation universals. In particular, the experiments on validating the universals of simplification, convergence and transfer will be detailed.

Following from this study, the speaker will outline the work on the use of comparable corpora to track language change over time, in particular the recent changes in lexical density and lexical richness in two consecutive thirty-year time periods in British English (1931–1961 and 1961–1991) and in American English from the 1960s to the 1990s (1961–1992).

Finally, the speaker will share the latest results from his work with colleagues on the use of comparable corpora for extracting and translating multiword expressions. The methodology developed does not rely on any dictionaries or parallel corpora, nor does it use any (bilingual) grammars. The only information comes from comparable corpora, inexpensively compiled with the help of the ACCURAT toolkit (Su and Babych 2012a) where only documents above a specific threshold were considered for inclusion. The presentation will conclude with the results of an interesting experiment as part of this study which sought to establish whether large loosely comparable data would yield better results than smaller but strictly comparable corpora.

Session 2: Building Comparable Corpora

Monday 23 May, 10:30 – 12:30

Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints

Andrey Kutuzov, Mikhail Kopotev, Tatyana Sviridenko and Lyubov Ivanova

We present our experience in applying distributional semantics (neural word embeddings) to the problem of representing and clustering documents in a bilingual comparable corpus. Our data is a collection of Russian and Ukrainian academic texts, for which topics are their academic fields. In order to build language-independent semantic representations of these documents, we train neural distributional models on monolingual corpora and learn the optimal linear transformation of vectors from one language to another. The resulting vectors are then used to produce 'semantic fingerprints' of documents, serving as input to a clustering algorithm.

The presented method is compared to several baselines including 'orthographic translation' with Levenshtein edit distance and outperforms them by a large margin. We also show that language-independent 'semantic fingerprints' are superior to multi-lingual clustering algorithms proposed in the previous work, at the same time requiring less linguistic resources.

A 2D CRF Model for Sentence Alignment

Yong Xu and François Yvon

The identification of parallel segments from parallel or comparable corpora can be performed at various levels. Alignments at the sentence level are useful for many downstream tasks, and also simplify the identification of finer grain correspondences. Most state-of-the-art sentence aligners are unsupervised, and attempt to infer endogenous alignment clues based on the analysis of the sole bitext. In decoding, they typically make simplifying assumptions, so that efficient dynamic programming techniques can be applied. Owing to such assumptions, high-precision sentence alignment remains difficult for certain types of corpora, in particular for literary texts. In this paper, we propose to learn a supervised alignment model, which represents the alignment matrix as two-dimensional Conditional Random Fields (2D CRF), converting sentence alignment into a structured prediction problem. This formalism enables us to take advantage of a rich set of overlapping features. Furthermore, it also allows us to relax some assumptions in decoding.

Parallel Document Identification using Zipf's Law

Mehdi Mohammadi

Parallel texts are an essential resource in many NLP tasks. One main issue to take advantage of these resources is to distinguish parallel or comparable documents that may have parallel fragments of texts from those that have no corresponding text. In this paper we propose a simple and efficient method to identify parallel documents based on Zipfian frequency distribution of available parallel corpora. In our method, we introduce a score called Cumulative Frequency Log by which we can measure the similarity of two documents that fit into a simple linear regression model. The regression model is generated based on the word ranks and frequencies of an available parallel corpus. The evaluation of the proposed approach over three language pairs achieve accuracy up to 0.86.

Session 3: Invited Presentation

Monday 23 May, 14.00–15.00

Exploring the Richness and Limitations of Web Sources for Comparable Corpus Research

Gregory Grefenstette

Comparable Corpora have been used to improve statistical machine translation, for augmenting linked open data, for finding terminology equivalents, and to create other linguistic resources for natural language processing and language learning applications. Recently, continuous vector space models, creating and exploiting word embeddings, have been gaining in popularity in more powerful solutions to creating, and sometimes replacing, these resources. Both classical comparable corpora solutions and vector space models require the presence of a large quantity of multilingual content. In this talk, we will discuss the breadth of this content on the internet to provide some type of intuition in how successful comparable corpus approaches will be in achieving its goals of providing multilingual and cross lingual resources. We examine current estimates of language presence and growth on the web, and of the availability of the type of resources needed to continue and extend comparable corpus research.

Session 4: Applications of Comparable Corpora

Monday 23 May, 15:00 – 16:00

A Mutual Iterative Enhancement Model for Simultaneous Comparable Corpora and Bilingual Lexicons Construction

Zede Zhu, Xinhua Zeng, Shouguo Zheng, Xiongwei Sun, Shaoqi Wang and Shizhuang Weng

Constructing bilingual lexicons from comparable corpora has been investigated in a two-stage process: building comparable corpora and mining bilingual lexicons, respectively. However, there are two potential challenges remaining, which are out-of-vocabulary words and different comparability degrees of corpora. To solve above problems, a novel iterative enhancement model is proposed for constructing comparable corpora and bilingual lexicons simultaneously under the assumption that both processes can be mutually reinforced. As compared to separate process, it is concluded that both simultaneous processes show better performance on different domain data sets via a small-volume general bilingual seed dictionary.

Hard Synonymy and Applications in Automatic Detection of Synonyms and Machine Translation

Ana Sabina Uban

We investigate in this paper the property of hard synonymy, defined as synonymy which is maintained across two or more languages. We use synonym dictionaries for four languages, as well as parallel corpora, and tools for distributional synonym extraction, in order to perform experiments to investigate the potential applications of hard synonymy for the automatic detection of synonyms and for machine translation. We show that hard synonymy can be used to discriminate between distributionally similar words that are true synonyms and those that are merely semantically related or even antonyms. We also investigate whether hard synonym word-translation pairs can be useful for lexical machine translation, by analyzing their occurrences in word-aligned parallel corpora. We build a database of words, synonyms and their translations for the four languages, including a generally low resourced language (Romanian) and show how it can be used to investigate properties of words and their synonyms cross-lingually.

Session 5: Discussion

Monday 23 May, 16:30 – 17:30

Towards Preparation of the Second BUCC Shared Task: Detecting Parallel Sentences in Comparable Corpora

Pierre Zweigenbaum, Serge Sharoff and Reinhard Rapp

In this paper we provide a summary of the rationale and the dataset contributing to the second shared task of the BUCC workshop. The shared task is aimed at detecting the best candidates for parallel sentences in a large text collection. The dataset for the shared task is based on a careful mix of parallel and non-parallel corpora. It contains 1.4 million French sentences and 1.9 million English sentences, in which 17 thousand sentence pairs are known to be parallel. The shared task itself is scheduled for the 2017 edition of the workshop.

LREC 2016 Workshop

CCURL 2016

Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity

23 May 2016

ABSTRACTS

Editors

Claudia Soria, Laurette Pretorius, Thierry Declerck, Joseph Mariani, Kevin Scannell, Eveline Wandl-Vogt

Workshop Programme

	Opening Session
09.15 – 09.30	Introduction
09.30 – 10.30	Jon French, *Oxford Global Languages: a Defining Project (Invited Talk)*
10.30 – 11.00	Coffee Break
	Session 1
11.00 – 11.25	Antti Arppe, Jordan Lachler, Trond Trosterud, Lene Antonsen, and Sjur N. Moshagen, *Basic Language Resource Kits for Endangered Languages: A Case Study of Plains Cree*
11.25 – 11.50	George Dueñas and Diego Gómez, *Building Bilingual Dictionaries for Minority and Endangered Languages with Mediawiki*
11.50 – 12.15	Dorothee Beermann, Tormod Haugland, Lars Hellan, Uwe Quasthoff, Thomas Eckart, and Christoph Kuras, *Quantitative and Qualitative Analysis in the Work with African Languages*
12.15 – 12.40	Nikki Adams and Michael Maxwell, *Somali Spelling Corrector and Morphological Analyzer*
12.40 – 14.00	Lunch Break
	Session 2
14.00 – 14.25	Delyth Prys, Mared Roberts, and Gruffudd Prys, *Reprinting Scholarly Works as e-Books for Under-Resourced Languages*
14.25 – 14.50	Cat Kutay, *Supporting Language Teaching Online*
14.50 – 15.15	Maik Gibson, *Assessing Digital Vitality: Analytical and Activist Approaches*
15.15 – 15.40	Martin Benjamin, *Digital Language Diversity: Seeking the Value Proposition*
15.40 – 16.00	Discussion
16.05 – 16.30	Coffee Break
16.30 – 17.30	**Poster Session**
	Sebastian Stüker, Gilles Adda, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Elodie Gauthier, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noel Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Markus Müller, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian, *Innovative Technologies for Under-Resourced Language Documentation: The BULB Project*
	Dirk Goldhahn, Maciej Sumalvico, and Uwe Quasthoff, *Corpus Collection for Under-Resourced Languages with More than One Million Speakers*
	Dewi Bryn Jones and Sarah Cooper, *Building Intelligent Digital Assistants for Speakers of a Lesser-Resourced Language*
	Justina Mandravickaite and Michael Oakes, *Multiword Expressions for Capturing Stylistic Variation Between Genders in the Lithuanian Parliament*
	Richard Littauer and Hugh Paterson III, *Open Source Code Serving Endangered Languages*
	Uwe Quasthoff, Dirk Goldhahn, and Sonja Bosch, *Morphology Learning for Zulu*
17.30 – 18.00	Discussion and Conclusions

Workshop Organizers

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

Workshop Programme Committee

Gilles Adda	LIMSI-CNRS & IMMI, France
Tunde Adegbola	African Languages Technology Initiative, Nigeria
Eduardo Avila	Rising Voices, Bolivia
Martin Benjamin	The Kamusi Project, Switzerland
Delphine Bernhard	LiLPa, Université de Strasbourg, LiLPA, France
Paul Bilbao Sarria	Euskararen Gizarte Erakundeen KONTSEILUA, Spain
Vicent Climent Ferrando	NPLD, Belgium
Daniel Cunliffe	Prifysgol De Cymru / University of South Wales, School of Computing and Mathematics, UK
Nicole Dolowy-Rybinska	Polska Akademia Nauk / Polish Academy of Sciences, Poland
Mikel Forcada	Universitat d'Alacant, Spain
Maik Gibson	SIL International, UK
Tjerd de Graaf	De Fryske Akademy, The Netherlands
Thibault Grouas	Délégation Générale à la langue française et aux langues de France, France
Auður Hauksdóttir	Vigdís Finnbogadóttir Institute of Foreign Languages, Iceland
Peter Juel Henrichsen	Copenhagen Business School, Denmark
Davyth Hicks	ELEN, France
Kristiina Jokinen	Helsingin Yliopisto / University of Helsinki, Finland
John Judge	ADAPT Centre, Dublin City University, Ireland
Steven Krauwer	CLARIN, The Netherlands
Steven Moran	Universität Zürich, Switzerland
Silvia Pareti	Google Inc., Switzerland
Daniel Pimienta	MAAYA
Steve Renals	University of Edinburgh, UK
Kepa Sarasola Gabiola	Euskal Herriko Unibertsitatea / University of the Basque Country, Spain
Felix Sasaki	DFKI GmbH and W3C fellow, Germany
Virach Sornlertlamvanich	Sirindhorn International Institute of Technology / Thammasat University, Thailand
Ferran Suay	Universitat de València, Spain
Jörg Tiedemann	Uppsala Universitet, Sweden
Francis M. Tyers	Norges Arktiske Universitet, Norway

Preface

The LREC 2016 Workshop on "Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity" (CCURL 2016) explores the relationship between language and the Internet, and specifically the web of documents and the web of data, as well as the emerging Internet of things, is a growing area of research, development, innovation and policy interest. The emerging picture is one where language profoundly affects a person's experience of the Internet by determining the amount of accessible information and the range of services that can be available, e.g. by shaping the results of a search engine, and the amount of everyday tasks that can be carried out virtually. The extent to which a language can be used over the Internet or in the Web not only affects a person's experience and choice of opportunities; it also affects the language itself.

If a language is poorly or not sufficiently supported to be used over digital devices, for instance if the keyboard of the device is not equipped with the characters and diacritics necessary to write in the language, or if there is no spell checker for a language, then its usability becomes severely affected, and it might never be used online. The language could become "digitally endangered", and its value and profile could be lessened, especially in the eyes of new generations. On the other hand, concerted efforts to develop a language technologically could contribute to the digital ascent and digital vitality of a language, and therefore to digital language diversity. These considerations call for a closer examination of a number of related issues.

First, the issue of "digital language diversity": the Internet appears to be far from linguistically diverse. With a handful of languages dominating the Web, there is a linguistic divide that parallels and reinforces the digital divide. The amount of information and services that are available in digitally less widely used languages are reduced, thus creating inequality in the digital opportunities and linguistic rights of citizens. This may ultimately lead to unequal digital dignity, i.e. uneven perception of a language importance as a function of its presence on digital media, and unequal opportunities for digital language survival.

Second, it is important to reflect on the conditions that make it possible for a language to be used over digital devices, and about what can be done in order to grant this possibility to languages other than the so-called "major" ones. Despite its increasing penetration in daily applications, language technology is still under development for these major languages, and with the current pace of technological development, there is a serious risk that some languages will be left wanting in terms of advanced technological solutions such as smart personal assistants, adaptive interfaces, or speech-to-speech translations. We refer to such languages as under-resourced. The notion of digital language diversity may therefore be interpreted as a digital universe that allows the comprehensive use of as many languages as possible.

All the papers accepted for the Workshop address at least one of these issues, thereby making a noteworthy contribution to the relevant scholarly literature and to the technological development of a wide variety of under-resourced languages. Each of the fifteen accepted papers was reviewed by at least three members of the Programme Committee, eight of which are presented as oral presentations and six as posters. We look forward to collaboratively and computationally building on this growing tradition of CCURL in the future for the continued benefit of all the under-resourced languages of the world!

C. Soria, L. Pretorius, T. Declerck, J. Mariani, K. Scannell, E. Wandl-Vogt May 2016

Invited Talk
Monday, 23 May 2016, 09:30–10:30

Oxford Global Languages: a Defining Project

Jon French

Oxford Global Languages is a major new initiative which will enable millions of people across the globe to find answers online to their everyday language questions in 100 of the world's languages. For the first time, large quantities of quality lexical information for these languages will be systematically created, collected, and made available, in a single linked repository, to speakers and learners. The OGL programme will:

- Enable the development of new digital tools and resources to revitalise and support under-represented world languages.
- Give these languages a living, growing, vibrant presence in the digital landscape.
- Document and include living languages including their variants and dialects, truly recording how they are used today.
- Provide an interactive community in which people can suggest new content, ask questions, and discuss language–these are living dictionaries of real languages that the community will help to build.

Which languages will be included? This ambitious initiative includes major global languages and digitally under-represented ones – those which are actively spoken and used by large communities but which have little digital capacity or accessibility. These digitally under-represented languages and their speakers are increasingly disadvantaged in social, business, and cultural areas of life because resource in the digital world is focused on a small number of globally predominant languages. Oxford Global Languages launched its first two language sites, isiZulu and Northern Sotho, in September 2015. Many more will be added over the next few years. See the first sites for yourself: isiZulu zu.oxforddictionaries.com Northern Sotho nso.oxforddictionaries.com

Session 1
Monday 23 May, 11:00–12:40

Basic Language Resource Kits for Endangered Languages: A Case Study of Plains Cree

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

Using Plains Cree as an example case, we describe and motivate the adaptation of the BLARK approach for endangered, less-resourced languages (resulting in an EL-BLARK), based on (1) what linguistic resources are most likely to be readily available, (2) which end-user applications would be of most practical benefit to these language communities, and (3) which computational linguistic technologies would provide the most reliable benefit with respect to the development efforts required.

Building Bilingual Dictionaries for Minority and Endangered Languages with Mediawiki

George Dueñas and Diego Gómez

In this paper, we present the ongoing work of the indigenous languages team at Caro and Cuervo Institute in developing language resources and technologies to document and revitalize minority languages which are in some degree of endangerment. This work consists in creating not only electronic dictionaries, but also a space where linguistic and cultural information is stored about the language as personal names and toponyms, among others. In order to do this, we have created different templates for placing the information. The main software that we have used to create these dictionaries is MediaWiki and the Semantic Mediawiki extension (free software open source). The Mediawiki software, adapted to lexicography needs, has become an important tool in this project. When the information has been stored in variables, we can display it through queries in the Semantic Mediawiki syntax. All of these tools have enabled us to show and recover information from each lexicographic entry. Two bilingual (uni- and bidirectional) dictionaries of Saliba and Carijona indigenous languages of Colombia were built with these tools. These dictionaries will increase the amount of content available for such languages in Internet and reduce a little the lack of places for using their languages over digital media.

Quantitative and Qualitative Analysis in the Work with African Languages

Dorothee Beermann, Tormod Haugland, Lars Hellan, Uwe Quasthoff, Thomas Eckart, and Christoph Kuras

We discuss the development of a combined search environment for the Leipzig Corpora Collection (LCC) and the TypeCraft Interlinear Glossed Text Repository (TC). This digital infrastructure facilitates corpus methodologies for under-resourced languages. By providing multiple-site accessibility of all data, we hope to give a new impetus to linguists and language experts to employ digital services for data analytics. The definition of export and import APIs using Web Services are shown to be useful for a collaboration between two different projects, and to extend and combine existing linguistic material. In this way we also increase the access to data from under-resourced languages.

Somali Spelling Corrector and Morphological Analyzer

Nikki Adams and Michael Maxwell

For any language, a basic requirement for a natural language processing capability is the availability of an electronic dictionary in a form which other NLP tools can use. For all but isolating (or nearly isolating) languages, another basic requirement is the capability to both generate and analyze all inflected forms. This second requirement is usually fulfilled by a finite state transducer that uses the morphological (and perhaps phonological) rules of the written language, together with the dictionary. A third need, for languages where there is significant variation in spelling, is a spell corrector, which can also be implemented as a finite state transducer. These three resources are mutually supportive: the morphological transducer requires the dictionary, and because of the properties of finite state grammars, it is simple to couple finite state transducers together, giving inflectional lookup of misspelled forms. And testing the parser and spell corrector on a web corpus can supply new words for the dictionary, completing the cycle. This paper reports on the cleaning of an electronic dictionary for Somali, the construction of a Somali morphological analyzer and a spelling corrector, and their resulting composed form. Somali is an Afro-Asiatic language in the Cushitic sub-family with complex morphology, complex morpho-phonological rules, and an orthography which, though officially standardized, is often not used consistently among speakers of the language. The electronic dictionary is showing signs of age; in particular, we believe there is need for expansion of its vocabulary to cover modern forms. While we have not as yet implemented dictionary expansion for Somali, we describe similar work in Yemeni and Sudanese Arabic, which could be extended to Somali.

Session 2
Monday, 23 May 2016, 14:00–16:00

Reprinting Scholarly Works as e-Books for Under-Resourced Languages

Delyth Prys, Mared Roberts, and Gruffudd Prys

This paper on the DECHE project for Digitization, E-publishing, and Electronic Corpus reports on a project undertaken for the Coleg Cymraeg Cenedlaethol, the virtual Welsh-medium College for universities in Wales. The DECHE project's aim is to digitize out of print scholarly works across multiple disciplines and help create a library of e-books available to Welsh-speaking academics and students. The context of e-book publication in Wales, and the digitization and e-book production process is described, together with the software tools used. The criteria for selecting a shortlist of books for inclusion are given, as are the types of books chosen. Attitudes and take-up of the students surveyed are also discussed, as are the dissemination of the resulting e-books, and statistics of use. This takes place in the context of the increasing popularity of e-books for education, including at university level, and their value for less-resourced language communities because of the lower production and distribution costs, and their contribution to raising the image and status of those languages as fit for purpose in a digital age.

Supporting Language Teaching Online

Cat Kutay

This paper presents the work done in a project to support the teaching of New South Wales (NSW) Indigenous languages through online and mobile systems. The process has incorporated languages with a variety of resources available and involved community workshops to engage speakers and linguists in developing and sharing these resources with learners. This research looks at Human Computer Interaction (HCI) for developing interfaces for Indigenous language learning, by considering the knowledge sharing practises used in the communities, and we compare this work in Australia with similar findings on language reclamation with the Penan indigenous people Malaysia. The HCI studies have been conducted in workshops with linguists and community members interested in studying and teaching their language. The web services developed and used for various languages uses processes of tacit knowledge sharing in an online environment.

Assessing Digital Vitality: Analytical and Activist Approaches

Maik Gibson

The digital vitality of a language is of concern to many, including linguists and members of the communities that use the language. Kornai (2013) has established a framework for measuring digital vitality which he has used to map the numbers of languages at four different levels - Digitally thriving, Vital, Heritage and Still. This overview is very useful in understanding the overall challenge that minority languages face in digital use. However, when working with a particular community which speaks a minority language, we argue that it is useful to add two more levels of analysis, which would be difficult to justify in an empirical study using mass comparison. As activists we need to be able to identify Emergent use which may not be visible to web crawling, for example in private messaging. Furthermore, identifying where conditions exist for possible digital ascent (e.g. literacy practices, intergenerational transmission of the language in the home) justifies a level we name Latent, which is by its very nature not something to be observed empirically. However the potential for digital development of a language at this level, is, we argue, different from one which is truly Still, unlikely to ascend. The inclusion of these categories might assist digital language activists and communities in effective planning for digital ascent, while not being useful categories for empirical mass comparison. Therefore we propose that action research with a linguistic community will benefit from a five-level framework, and demonstrate how it may be used.

Digital Language Diversity: Seeking the Value Proposition

Martin Benjamin

This paper is a response to the CCURL workshop call for discussion about issues pertaining to the creation of an Alliance for Digital Language Diversity. As a global project, Kamusi has been building collaborative relationships with numerous organizations, becoming more familiar than most with global activities and the global funding situation for less-resourced languages. This paper reviews the experiences of many involved with creating or using digital resources for diverse languages, with an analysis of who finds such resources important, who does not, what brings such resources into existence, and what the barriers are to the wider development of inclusive language technology. It is seen that practitioners face obstacles to maximizing the effects of their own work and gaining from the advances of others due to a funding environment that does not recognize the value of linguistic resources for diverse languages, as either a social or economic good. Proposed solutions include the normalization of the expectation that digital services will be available in major local languages, international legal requirements for language provision on par with European regulations, involvement of speaker communities in the guided production of open linguistic resources, and the formation of a research consortium that can together build a common linguistic data infrastructure.

Poster Session
Monday, 23 May 2016, 16:30–17:30

Innovative Technologies for Under-Resourced Language Documentation: The BULB Project

Sebastian Stüker, Gilles Adda, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Elodie Gauthier, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noel Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Markus Müller, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian

The project Breaking the Unwritten Language Barrier (BULB), which brings together linguists and computer scientists, aims at supporting linguists in documenting unwritten languages. To achieve this, we develop tools tailored to the needs of documentary linguists by building upon technology and expertise from the area of natural language processing, most prominently automatic speech recognition and machine translation. As a development and test bed for this, we have chosen three less-resourced African languages from the Bantu family: Basaa, Myene and Embosi. Work within the project is divided into three main steps: 1) Collection of a large corpus of speech (100h per language) at a reasonable cost. After initial recording, the data is re-spoken by a reference speaker to enhance the signal quality and orally translated into French. 2) Automatic transcription of the Bantu languages at phoneme level and the French translation at word level. The recognized Bantu phonemes and French words will then be automatically aligned to extract Bantu's morphemes. 3) Tool development. In close cooperation and discussion with the linguists, the speech and language technologists will design and implement tools that will support the linguists in their work, taking into account the linguists' needs and technology capabilities. The data collection has begun for the three languages. We have been using standard mobile devices and a dedicated software— LIG-AIKUMA, which offers a range of speech collection modes (recording, re-speaking, translation and elicitation). LIG-AIKUMA's improved features include a smart generation and handling of speaker metadata as well as re-speaking and parallel audio data mapping.

Corpus Collection for Under-Resourced Languages with More than One Million Speakers

Dirk Goldhahn, Maciej Sumalvico, and Uwe Quasthoff

For only 40 of about 350 languages with more than one million speakers, the situation concerning text resources is comfortable. For the remaining languages, the number of speakers indicates a need for both corpora and tools. This paper describes a corpus collection initiative for these languages. While random Web crawling has serious limitations, native speakers with knowledge of web pages in their language are of invaluable help. The aim is to give interested scholars, language enthusiasts the possibility to contribute to corpus creation or extension by simply entering a URL into the Web Interface. Using a Web portal URLs of interest are collected with the help of the respective communities. A standardized corpus processing chain for daily newspaper corpora creation is adapted to append newly added web pages to an increasing corpus. As a result we will be able to collect larger corpora for under-resourced languages by a community effort. These corpora will be made publicly available.

Building Intelligent Digital Assistants for Speakers of a Lesser-Resourced Language

Dewi Bryn Jones and Sarah Cooper

This paper reports on the work to develop intelligent digital assistants for speakers of a lesser-resourced language, namely Welsh. Such assistants provided by commercial vendors such as Apple (Siri), Amazon (Alexa), Microsoft (Cortana) and Google (Google Now) are allowing users increasingly to speak in natural English with their devices and computers in order to complete tasks, obtain assistance and request information. We demonstrate how these systems' architectures do not provide the means for external developers to build intelligent speech interfaces for additional languages, which, in the case of less resourced languages, are likely to remain unsupported. Consequently we document how such an obstacle has been tackled with open alternatives. The paper highlights how previous work on Welsh language speech recognition were improved, utilized and integrated into an existing open source intelligent digital assistant software project. The paper discusses how this work hopes to stimulate further developments and include Welsh and other lesser-resourced languages in as many developments of intelligent digital assistants as possible.

Multiword Expressions for Capturing Stylistic Variation Between Genders in the Lithuanian Parliament

Justina Mandravickaite and Michael Oakes

The relation between gender and language has been studied by many authors, but there is no general agreement regarding gender influence on language usage in the professional environment. This could be because in most of the studies data sets are too small or texts of individual authors are too short in order to capture differences of language usage according to gender successfully. This study draws on a larger corpus of transcribed speeches in the Lithuanian Parliament (1990-2013) to explore gender differences in a language with a setting of political debates using stylometric analysis. The experimental set up consists of multiword expressions as features (formulaic language can allow a more detailed interpretation of the results in comparison to character n-grams or even most frequent words) combined with unsupervised machine learning algorithms to avoid the class imbalance problem. MWEs as features in combination with distance measures and hierarchical clustering were successful in capturing and mapping difference in speech according to gender in the Lithuanian Parliament. Our results agree with the experimental outcomes of Hoover (2002) and Hoover (2003), where frequent word sequences and collocations combined with clustering showed more accurate results than just frequent words.

Open Source Code Serving Endangered Languages

Richard Littauer and Hugh Paterson III

We present a database of open source code that can be used by low-resource language communities and developers to build digital resources. Our database is also useful to software developers working with those communities and to researchers looking to describe the state of the field when seeking funding for development projects.

Morphology Learning for Zulu

Uwe Quasthoff, Dirk Goldhahn, and Sonja Bosch

Morphology is known to follow structural regularities, but there are always exceptions. The number and complexity of the exceptions depend on the language under consideration. Substring classifiers are shown to perform well for different languages with different amounts of training data. For a less resourced Bantu language like Zulu the learning curve is compared to that of a well-resourced language like German, the learning curve of which might be considered as an extrapolation for the less resourced languages in the case of larger training set sizes. Two substring classifiers, TiMBL and the Compact Patricia Tree Classifier are tested and shown to give comparable results.

LREC 2016 Workshop: Abstracts
"CCURL 2016 – Collaboration and Computing for Under-Resourced Languages: Towards an Alliance for Digital Language Diversity"

23 May 2016 – Portorož, Slovenia

Edited by Claudia Soria, Laurette Pretorius, Thierry Declerck, Joseph Mariani, Kevin Scannell, Eveline Wandl-Vogt

`http://www.ilc.cnr.it/ccurl2016/`

Acknowledgments: the CCURL 2016 Workshop is endorsed by the Erasmus+ DLDP project (grant agreement no.: 2015-1-IT02-KA204-015090).

the Digital Language Diversity Project

Improving Social Inclusion Using NLP: Tools and Resources

23 May 2016

ABSTRACTS

Editors:

Ineke Schuurman, Vincent Vandeghinste, Horacio Saggion

Workshop Programme

09:00 – 09:15 Opening

09:15 – 09:40
Liz Tilly, *Issues relating to using a co-productive approach in an accessible technology project*

09:40 – 10:05
Krzysztof Wróbel, Dawid Smoleń, Dorota Szulc and Jakub Gałka, *Development of the First Polish Sign Language Part-of-Speech Tagger*

10:05 – 10:30
Leen Sevens, Tom Vanallemeersch, Ineke Schuurman, Vincent Vandeghinste and Frank Van Eynde, *Automated Spelling Correction for Dutch Internet Users with Intellectual Disabilities*

10:30 – 11:00 Coffee break

11:00 – 11:25
Victoria Yaneva, Richard Evans and Irina Temnikova, *Predicting Reading Difficulty for Readers with Autism Spectrum Disorder*

11:25 – 11:50
Estela Saquete, Ruben Izquierdo Bevia and Sonia Vazquez, *SimplexEduReading: Simplification of Natural Language for Reading Comprehension Improvement in Education*

11:50 – 12:40
Lucia Specia (invited speaker), *Text Simplification for Social Inclusion*

12:40 – 12:55 General discussion

12:55 – 13:00 Closing

Workshop Organizers

Ineke Schuurman	KU Leuven (BE)
Vincent Vandeghinste	KU Leuven (BE)
Horacio Saggion	Universitat Pompeu Fabra, Barcelona (ES)

Workshop Programme Committee

Susana Bautista	Federal University of Rio Grande do Sul (BR)
Heidi Christensen	University of Sheffield (UK)
Onno Crasborn	Radboud University (NL)
Koenraad De Smedt	University of Bergen (NO)
Nuria Gala	Aix-Marseille Université (FR)
Peter Ljunglöf	University of Gothenburg (SE)
Isa Maks	VU University Amsterdam (NL)
Davy Nijs	UC Leuven-Limburg (BE)
Jean-Pierre Martens	Ghent University (BE)
Martin Reynaert	Tilburg University and Radboud University (NL)
Horacio Saggion	Universitat Pompeu Fabra (ES)
Ineke Schuurman	University of Leuven (BE)
Liz Tilly	University of Wolverhampton (UK)
Vincent Vandeghinste	University of Leuven (BE)
Hugo Van hamme	University of Leuven (BE)

Introduction

Social media are an inherent part of life in the 21st century and should be accessible to anyone. People who are to some extent functionally illiterate are currently excluded from properly using social media such as Twitter, Facebook, WhatsApp.
Therefore tools and resources are needed that aim to bridge the social divide by providing technology that allows more or less functionally illiterate users to be included in the current and future textual social media.
Such technologies and resources allow or facilitate the conversion of non-verbal input (like signs, pictograms) into textual output, or allow to improve ill-formed input into proper text. In the other direction, the technologies and resources allow to automatically augment textual communication by other media, such as pictures, signs, video, and other forms of Augmentative and Alternative Communication, or to simplify and reduce messages such that the information is disclosed to the targeted users, and the users are no longer excluded from taking part in social media.

In the following some tools and/or resources are presented.

Issues relating to using a co-productive approach in an accessible technology project

Liz Tilly; Faculty of Education, Health and Wellbeing, University of Wolverhampton, United Kingdom

This paper discusses the issue of accessibility to being online and using social media for people with a learning disability, and the challenges to using a co-production approach in an accessible technology project. While an increasing number of daily living tasks are now completed online, people with a learning disability frequently experience digital exclusion due to limited literacy and IT skills. The Able to Include project sought to engage people with a learning disability as active partners to test and feedback on the use and development of a pictogram app used to make social media more accessible. The challenges mainly related to the feedback needing to be sent electronically to the partners; there was only minimal contact with them and no face to face contact. The paper also outlines how other challenges were overcome to enable genuine and meaningful co-production. These included addressing online safety and ethical issues regarding anonymity.

Development of the First Polish Sign Language Part-of-Speech Tagger

Krzysztof Wróbel, Dawid Smoleń, Dorota Szulc and Jakub Gałka; Department of Computational Linguistics, Jagiellonian University, Krakow, Poland

This article presents the development of the first part-of-speech (POS) tagger for Polish Sign Language (PJM). Due to the lack of PJM corpora, a data set consisting of 34.5 thousand sentences was automatically created and annotated. It was done using a machine translation (MT) system, from Polish to PJM. The annotation with POS tags is done concurrently by transferring and mapping them from Polish. The POS tagger is trained using a sequence classifier and tested on a manually-developed PJM corpus. The results are compared to other taggers for various languages, and error analysis is performed. This paper shows that it is possible to develop a POS tagger with promising results using a transfer-based MT system. The created PJM corpus will be publicly shared.

Automated Spelling Correction for Dutch Internet Users with Intellectual Disabilities

Leen Sevens, Tom Vanallemeersch, Ineke Schuurman, Vincent Vandeghinste and Frank Van Eynde; Centre for Computational Linguistics, Kniversity of Leuven, Belgium

We present the first version of an automated spelling correction system for Dutch Internet users with Intellectual Disabilities (ID). The normalization of ill-formed messages is an important preprocessing step before any conventional Natural Language Processing (NLP) process can be applied. As such, we describe the effects of automated correction of Dutch ID text within the larger framework of a Text-to-Pictograph translation system. The present study consists of two main parts. First, we thoroughly analyze email messages that have been written by users with cognitive disabilities in order to gain insights on how to develop solutions that are specifically tailored to their needs. We then present a new, generally applicable approach toward context-sensitive spelling correction, based on character-level fuzzy matching techniques. The resulting system shows significant improvements, although further research is still needed.

Predicting Reading Difficulty for Readers with Autism Spectrum Disorder
Victoria Yaneva, Richard Evans* and Irina Temnikova**, *Research Institute in Information and Language Processing, University of Wolverhampton, UK and **Qatar Computing Research Institute, HBKU, Doha, Qatar*

People with autism experience various reading comprehension difficulties, which is one explanation for the early school dropout, reduced academic achievement and lower levels of employment in this population. To overcome this issue, content developers who want to make their textbooks, websites or social media accessible to people with autism (and thus for every other user) but who are not necessarily experts in autism, can benefit from tools which are easy to use, which can assess the accessibility of their content, and which are sensitive to the difficulties that autistic people might have when processing texts/websites. In this paper we present a preliminary machine learning readability model for English developed specifically for the needs of adults with autism. We evaluate the model on the ASD corpus, which has been developed specifically for this task and is, so far, the only corpus for which readability for people with autism has been evaluated.

SimplexEduReading: Simplification of Natural Language for Reading Comprehension Improvement in Education
Estela Saquete, Ruben Izquierdo Bevia** and Sonia Vazquez*; *University of Alicante, Spain and **Vrije Universiteit Amsterdam. The Netherlands*

The main aim of this paper is presenting a system, known as SimplexEduReading, capable of transforming educational natural language texts in Spanish into simpler and enriched texts in order to improve the reading comprehension process. The goal is to help people with comprehension problems, for instance, deaf people or people who are learning a language. The transformation process consists of applying different Natural Language Processing techniques in order to automatically detect linguistic features involved in the problem and: a) simplify the text preserving the original meaning, and b) enrich the text with very simple additional information. The transformations and supporting information that the system would provide include: 1) detecting name entities providing extra information about them, for example, related images, information from Wikipedia or synonyms; 2) detecting and resolving temporal expressions giving also a chronological timeline of the events; 3) simplifying complex sentences dividing them into simpler ones; 4) providing the definition of any word of the text; and 5) detecting the main topics of the text in order to easily provide a context of the text to the reader.

Text Simplification for Social Inclusion (Invited talk)
Lucia Specia; Natural Language Processing group, Department of Computer Science, University of Sheffield, United Kingdom

LREC 2016 Workshop

Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric Impairments (RaPID-2016)

Date: Monday, 23rd of May 2016

ABSTRACTS

Editor:

Dimitrios Kokkinakis

LREC 2016 Workshop: Abstracts
“Resources and Processing of Linguistic and Extra-Linguistic Data from People with Various Forms of Cognitive/Psychiatric impairments (RaPID-2016)”

23 May 2016 – Portorož, Slovenia

Edited by Dimitrios Kokkinakis

http://spraakbanken.gu.se/eng/rapid-2016

Acknowledgments: This work has received support from Riksbankens Jubileumsfond - The Swedish Foundation for Humanities and Social Sciences through the grant agreement no:NHS14-1761:1; the Centre for Ageing and Health (AGECAP) and the Speech and Language Processing for Assistive Technologies (SIG-SLPAT).

Workshop Programme

Time
14:00 – 14:10 – Welcome and Introduction

14:10-14:55 Invited keynote talk by Dr Peter Garrard, St George's, University of London:
Neurobehavioural disease signatures in language corpora

15:00-16:00 Session A (20+10 min x 2 papers)
Kathleen C. Fraser and Graeme Hirst, *Detecting semantic changes in Alzheimer's disease with vector space models*

Christine Howes, Mary Lavelle, Patrick G.T. Healey, Julian Hough and Rose McCabe: *Helping hands? Gesture and self-repair in schizophrenia*

16:00 – 16:30 Coffee break

16:30-17:50 Session B (15+5 min x 4 papers)
Spyridoula Varlokosta, Spyridoula Stamouli, Athanassios Karasimos, Georgios Markopoulos, Maria Kakavoulia, Michaela Nerantzini, Aikaterini Pantoula, Valantis Fyndanis, Alexandra Economou and Athanassios Protopapas: *A Greek Corpus of Aphasic Discourse: Collection, Transcription, and Annotation Specifications*

Mariya Khudyakova, Mira Bergelson, Yulia Akinina, Ekaterina Iskra, Svetlana Toldova and Olga Dragoy: *Russian CliPS: a Corpus of Narratives by Brain-Damaged Individuals*

Maksim Belousov, Mladen Dinev, Rohan M. Morris, Natalie Berry, Sandra Bucci and Goran Nenadic: *Mining Auditory Hallucinations from Unsolicited Twitter Posts*

Christopher Bull, Dommy Asfiandy, Ann Gledson, Joseph Mellor, Samuel Couth, Gemma Stringer, Paul Rayson, Alistair Sutcliffe, John Keane, Xiaojun Zeng, Alistair Burns, Iracema Leroi, Clive Ballard and Pete Sawyer: *Combining data mining and text mining for detection of early stage dementia: the SAMS framework*

17:50 – 18:00 Closing remarks

Workshop Organizers

Dimitrios Kokkinakis	University of Gothenburg, Sweden
Graeme Hirst	University of Toronto, Canada
Natalia Grabar	Université de Lille, France
Arto Nordlund	The Sahlgrenska Academy, Sweden
Jens Edlund	KTH - Royal Institute of Technology, Sweden
Åsa Wengelin	University of Gothenburg, Sweden
Simon Dobnik	University of Gothenburg, Sweden
Marcus Nyström	University of Lund, Sweden

Workshop Programme Committee

Jan Alexandersson	DFKI GmbH, Germany
Jonas Beskow	KTH - Royal Institute of Technology, Sweden
Heidi Christensen	University pf Sheffield, UK
Simon Dobnik	University of Gothenburg, Sweden
Jens Edlund	KTH - Royal Institute of Technology, Sweden
Gerardo Fernández	Universidad Nacional del Sur, Argentina
Peter Garrard	St George's, University of London, UK
Kallirroi Georgila	University of Southern California, USA
Natalia Grabar	Université de Lille, France
Nancy L. Green	U. of North Carolina at Greensboro, USA
Katarina Heimann Mühlenbock	University of Gothenburg, Sweden
Graeme Hirst	University of Toronto, Canada
Kristy Hollingshead	Florida Institute for Human & Machine Cognition (IHMC), USA
William Jarrold	Nuance Communications, USA
Richard Johansson	University of Gothenburg, Sweden
Dimitrios Kokkinakis	University of Gothenburg, Sweden
Yiannis Kompatsiaris	Centre for Research & Technology Hellas, Greece
Alexandra König	Toronto Rehabilitation Institute, Canada
Peter Ljunglöf	Chalmers University of Technology, Sweden
Karmele López-de-Ipiña	U. of the Basque Country (UPV/EHU), Spain
Arto Nordlund	The Sahlgrenska Academy, Sweden
Marcus Nyström	University of Lund, Sweden
François Portet	Laboratoire d'informatique de Grenoble, France
Vassiliki Rentoumi	SKEL, NCSR Demokritos, Greece
Frank Rudzicz	University of Toronto, Canada
Paul Thompson	Dartmouth College, USA
Magda Tsolaki	Aristotle University of Thessaloniki, Greece
Spyridoula Varlokosta	National & Kapodistrian U. of Athens, Greece
Åsa Wengelin	University of Gothenburg, Sweden
Maria Wolters	University of Edinburgh, UK

Preface/Introduction

The purpose of the Workshop on "*Resources and ProcessIng of linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric impairments*" (RaPID-2016) was to provide a snapshot view of some of the current technological landscape, resources, data samples and also needs and challenges in the area of processing various data from individuals with various types of mental and neurological health impairments and similar conditions at various stages; increase the knowledge, understanding, awareness and ability to achieve useful outcomes in this area and strengthen the collaboration between researchers and workers in the field of clinical/nursing/medical sciences and those in the field of language technology/computational linguistics/Natural Language Processing (NLP).

Although many of the causes of cognitive and neuropsychiatric impairments are difficult to foresee and accurately predict, physicians and clinicians work with a wide range of factors that potentially contribute to such impairments, e.g., traumatic brain injuries, genetic predispositions, side effects of medication, and congenital anomalies. In this context, there is new evidence that the acquisition and processing of linguistic data (e.g., spontaneous story telling) and extra-linguistic and production measures (e.g., eye tracking) could be used as a complement to clinical diagnosis and provide the foundation for future development of objective criteria to be used for identifying progressive decline or degeneration of normal mental and brain functioning.

An important new area of research in NLP emphasizes the processing, analysis, and interpretation of such data and current research in this field, based on linguistic-oriented analysis of text and speech produced by such a population and compared to healthy adults, has shown promising outcomes. This is manifested in early diagnosis and prediction of individuals at risk, the differentiation of individuals with various degrees of severity forms of brain and mental illness, and for the monitoring of the progression of such conditions through the diachronic analysis of language samples or other extra-linguistic measurements. Initially, work was based on written data but there is a rapidly growing body of research based on spoken samples and other modalities.

Nevertheless, there remains significant work to be done to arrive at more accurate estimates for prediction purposes in the future and more research is required in order to reliably complement the battery of medical and clinical examinations currently undertaken for the early diagnosis or monitoring of, e.g., neurodegenerative and other brain and mental disorders and accordingly, aid the development of new, non-invasive, time and cost-effective and objective (future) clinical tests in neurology, psychology, and psychiatry.

Papers were invited in all of the areas outlined in the *topics of interest* below particularly emphasizing multidisciplinary aspects of processing such data and also on the exploitation of results and outcomes and related ethical questions. Specifically, in the call for papers we solicited papers on the following topics:

- Building and adapting domain relevant linguistic resources, data, and tools, and making them available.
- Data collection methodologies.
- Acquisition of novel data samples, e.g. from digital pens (i.e., digital pen strokes) or keylogging and integrating them with data from various sources (i.e., information fusion).
- Guidelines, annotation schemas, and tools (e.g., for semantic annotation of data sets).

- Addressing the challenges of representation, including dealing with data sparsity and dimensionality issues, and feature combination from different sources and modalities,
- Adaptation of standard NLP tools to the domain.
- Syntactic, semantic, and pragmatic analysis of data, including modelling of perception (e.g., eye-movement measures of reading) and production processes (e.g., recording the writing process with digital pens, keystroke logging, etc.), use of gestures accompanying speech and non-linguistic behaviour.
- Machine learning approaches for early diagnosis, prediction, monitoring, classification, etc. of various cognitive, psychological, and psychiatric impairments, including unsupervised methods (e.g., distributional semantics).
- Evaluation of tools, systems, components, metrics, applications, and technologies that make use of NLP in the domain.
- Evaluation, comparison, and critical assessment of resources.
- Evaluation of the significance of extracted features.
- Involvement of medical professionals and patients and ethical questions.
- Deployment of resources.
- Experiences, lessons learned, and the future of NLP in the area.

Most of these topics lie at the heart of the papers that were accepted to the workshop which features 6 oral presentations.

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 also grateful to Dr Peter Garrard for accepting to give an invited talk at the workshop entitled: “Neurobehavioural disease signatures in language corpora”.

The Editor

Session A

Date / Time *Monday 23rd of May, 15:00 – 16:00*
Chairperson: Dimitrios Kokkinakis

Detecting semantic changes in Alzheimer's disease with vector space models

Kathleen C. Fraser and Graeme Hirst

Numerous studies have shown that language impairments, particularly semantic deficits, are evident in the narrative speech of people with Alzheimer’s disease from the earliest stages of the disease. Here, we present a novel technique for capturing those changes, by comparing distributed word representations constructed from healthy controls and Alzheimer’s patients. We investigate examples of words with different representations in the two spaces, and link the semantic and contextual differences to findings from the Alzheimer’s disease literature.

Helping hands? Gesture and self-repair in schizophrenia

Christine Howes, Mary Lavelle, Patrick G.T. Healey, Julian Hough and Rose McCabe

Successful social encounters require mutual understanding between interacting partners, and patients with schizophrenia are known to experience difficulties in social interaction. Several studies have shown that in general people compensate for verbal difficulties by employing additional multimodal resources such as hand gesture. We hypothesise that this will be impaired in patients with schizophrenia, and present a preliminary study to address this question. The results show that during social interaction, schizophrenia patients repair their own speech less. In addition, although increased hand gesture is correlated with increased self-repair in healthy controls, there is no such association in patients with schizophrenia, or their interlocutors. This suggests that multimodal impairments are not merely seen on an individual level but may be a feature of patients’ social encounters.

Session B

Date / Time *Monday 23rd of May, 16:30 – 17:50*
Chairperson: Mark Liberman

A Greek Corpus of Aphasic Discourse: Collection, Transcription, and Annotation Specifications

Spyridoula Varlokosta, Spyridoula Stamouli, Athanassios Karasimos, Georgios Markopoulos, Maria Kakavoulia, Michaela Nerantzini, Aikaterini Pantoula, Valantis Fyndanis, Alexandra Economou and Athanassios Protopapas

In this paper, the process of designing an annotated Greek Corpus of Aphasic Discourse (GREECAD) is presented. Given that resources of this kind are quite limited, a major aim of the GREECAD was to provide a set of specifications which could serve as a methodological basis for the development of other relevant corpora, and, therefore, to contribute to the future research in this area. The GREECAD was developed with the following requirements: a) to include a rather homogeneous sample of Greek as spoken by individuals with aphasia; b) to document speech samples with rich metadata, which include demographic information, as well as detailed information on the patients' medical record and neuropsychological evaluation; c) to provide annotated speech samples, which encode information at the micro-linguistic (words, POS, grammatical errors, clause types, etc.) and discourse level (narrative structure elements, main events, evaluation devices, etc.). In terms of the design of the GREECAD, the basic requirements regarding data collection, metadata, transcription, and annotation procedures were set. The discourse samples were transcribed and annotated with the ELAN tool. To ensure accurate and consistent annotation, a Transcription and Annotation Guide was compiled, which includes detailed guidelines regarding all aspects of the transcription and annotation procedure..

Russian CliPS: a Corpus of Narratives by Brain-Damaged Individuals

Mariya Khudyakova, Mira Bergelson, Yulia Akinina, Ekaterina Iskra, Svetlana Toldova and Olga Dragoy

In this paper we present a multimedia corpus of Pear film retellings by people with aphasia (PWA), right hemisphere damage (RHD), and healthy speakers of Russian. Discourse abilities of brain-damaged individuals are still under discussion, and Russian CliPS (Clinical Pear Stories) corpus was created for the thorough analysis of micro- and macro-linguistic levels of narratives by PWA and RHD. The current version of Russian CliPS contains 39 narratives by people with various forms of aphasia due to left hemisphere damage, 5 narratives by people with right hemisphere damage and no aphasia, and 22 narratives by neurologically healthy adults. The annotation scheme of Russian CliPS 1.0 includes the following tiers: quasiphonetic, lexical, lemma, part of speech tags, grammatical properties, errors, laughter, segmentation into clauses and utterances. Also analysis of such measures as informativeness, local and global coherence, anaphora, and macrostructure is planned as a next stage of the corpus development.

Mining Auditory Hallucinations from Unsolicited Twitter Posts

Maksim Belousov, Mladen Dinev, Rohan M. Morris, Natalie Berry, Sandra Bucci and Goran Nenadic

Auditory hallucinations are common in people who experience psychosis and psychotic-like phenomena. This exploratory study aimed to establish the feasibility of harvesting and mining datasets from unsolicited Twitter posts to identify potential auditory hallucinations. To this end, several search queries were defined to collect posts from Twitter. A training sample was annotated by research psychologists for relatedness to auditory hallucinatory experiences and a text classifier was trained on that dataset to identify tweets related to auditory hallucinations. A number of features were used including sentiment polarity and mentions of specific semantic classes, such as fear expressions, communication tools and abusive language. We then used the classification model to generate a dataset with potential mentions of auditory hallucinatory experiences. A preliminary analysis of a dataset (N = 4957) revealed that posts linked to auditory hallucinations were associated with negative sentiments. In addition, such tweets had a higher proportionate distribution between the hours of 11pm and 5am in comparison to other tweets.

Combining data mining and text mining for detection of early stage dementia: the SAMS framework

Christopher Bull, Dommy Asfiandy, Ann Gledson, Joseph Mellor, Samuel Couth, Gemma Stringer, Paul Rayson, Alistair Sutcliffe, John Keane, Xiaojun Zeng, Alistair Burns, Iracema Leroi, Clive Ballard and Pete Sawyer

In this paper, we describe the open-source SAMS framework whose novelty lies in bringing together both data collection (keystrokes, mouse movements, application pathways) and text collection (email, documents, diaries) and analysis methodologies. The aim of SAMS is to provide a non-invasive method for large scale collection, secure storage, retrieval and analysis of an individual's computer usage for the detection of cognitive decline, and to infer whether this decline is consistent with the early stages of dementia. The framework will allow evaluation and study by medical professionals in which data and textual features can be linked to deficits in cognitive domains that are characteristic of dementia. Having described requirements gathering and ethical concerns in previous papers, here we focus on the implementation of the data and text collection components.

Emotion and Sentiment Analysis

ABSTRACTS

Editors:

J. Fernando Sánchez-Rada, Björn Schuller

Workshop Programme

23 May 2016

9:00 – 9.10	Introduction by Workshop Chair
9.10 – 10:30	**Social Media**
Cristina Bosco et al.	Tweeting in the Debate about Catalan Elections
Ian D. Wood and Sebastian Ruder	Emoji as Emotion Tags for Tweets
Antoni Sobkowicz and Wojciech Stokowiec	Steam Review Dataset - new, large scale sentiment dataset
10:30 – 11:00	Coffee break
11:00 – 13:00	**Corpora and Data Collection**
Ebuka Ibeke et al.	A Curated Corpus for Sentiment-Topic Analysis
Jasy Liew Suet Yan and Howard R. Turtle	EmoCues-28: Extracting Words from Emotion Cues for a Fine-grained Emotion Lexicon
Lea Canales et al.	A Bootstrapping Technique to Annotate Emotional Corpora Automatically
Francis Bond et al.	A Multilingual Sentiment Corpus for Chinese, English and Japanese
13:00 – 14:00	Lunch break
14:00 – 15:00	**Personality and User Modelling**
Shivani Poddar et al.	PACMAN: Psycho and Computational Framework of an Individual (Man)
Veronika Vincze1, Klára Hegedűs, Gábor Berend and Richárd Farkas	Telltale Trips: Personality Traits in Travel Blogs
15:00 – 16:00	**Linked Data and Semantics**
Minsu Ko	Semantic Classification and Weight Matrices Derived from the Creation of Emotional Word Dictionary for Semantic Computing
J. Fernando Sánchez-Rada et al.	Towards a Common Linked Data Model for Sentiment and Emotion Analysis
16:00 – 16:30	Coffee break
16:30 – 18:00	**Beyond Text Analysis**
Bin Dong, Zixing Zhang and Björn Schuller	Empirical Mode Decomposition: A Data-Enrichment Perspective on Speech Emotion Recognition
Rebekah Wegener, Christian Kohlschein, Sabina Jeschke and Björn Schuller	Automatic Detection of Textual Triggers of Reader Emotion in Short Stories
Andrew Moore, Paul Rayson and Steven Young	Domain Adaptation using Stock Market Prices to Refine Sentiment Dictionaries

Workshop Organizers

J. Fernando Sánchez-Rada*	UPM, Spain
Björn Schuller *	Imperial College London, United Kingdom
Gabriela Vulcu	Insight Centre for Data Analytics, NUIG, Ireland
Carlos A. Iglesias	UPM, Spain
Paul Buitelaar	Insight Centre for Data Analytics, NUIG, Ireland
Laurence Devillers	LIMSI, France

Workshop Programme Committee

Elisabeth André	University of Augsburg, Germany
Noam Amir	Tel-Aviv University, Israel
Rodrigo Agerri	EHU, Spain
Cristina Bosco	University of Torino, Italy
Felix Burkhardt	Deutsche Telekom, Germany
Antonio Camurri	University of Genova, Italy
Montse Cuadros	VicomTech, Spain
Julien Epps	NICTA, Australia
Francesca Frontini	CNR, Italy
Diana Maynard	University of Sheffield, United Kingdom
Sapna Negi	Insight Centre for Data Analytics, NUIG, Ireland
Viviana Patti	University of Torino, Italy
Albert Salah	Boğaziçi University, Turkey
Jianhua Tao	CAS, P.R. China
Michel Valstar	University of Nottingham, United Kingdom
Benjamin Weiss	Technische Universität Berlin, Germany
Ian Wood	Insight Centre for Data Analytics, NUIG, Ireland

Preface/Introduction

ESA 2016 is the sixth edition of the highly successful series of Corpora for Research on Emotion. As its predecessors, the aim of this workshop is to connect the related fields around sentiment, emotion and social signals, exploring the state of the art in applications and resources. All this, with a special interest on multidisciplinarity, multilingualism and multimodality. This workshop is a much needed effort to fight the scarcity of quality annotated resources for emotion and sentiment research, especially for different modalities and languages.

This year's edition once again puts an emphasis on common models and formats, as a standardization process would foster the creation of interoperable resources. In particular, researchers have been encouraged to share their experience with Linked Data representation of emotions and sentiment, or any other application of Linked Data in the field, such as enriching existing data or publishing corpora and lexica in the Linked Open Data cloud.

Approaches on semi-automated and collaborative labeling of large data archives are also of interest, such as by efficient combinations of active learning and crowdsourcing, in particular for combined annotations of emotion, sentiment, and social signals. Multi- and cross-corpus studies (transfer learning, standardisation, corpus quality assessment, etc.) are further highly relevant, given their importance in order to test the generalisation power of models.

The workshop is supported by the Linked Data Models for Emotion and Sentiment Analysis W3C Community Group [1], the Association for the Advancement of Affective Computing [2] and the SSPNet [3] – some of the members of the organizing committee of the present workshop are executive members of these bodies.

As organising committee of this workshop, we would like to thank the organisers of LREC 2016 for their tireless efforts and for accepting ESA as a satellite workshop. We also thank every single member of the programme committee for their support since the announcement of the workshop, and their hard work with the reviews and feedback. Last, but not least, we are thankful to the community for the overwhelming interest and number of high-quality submissions. This is yet another proof that the emotion and sentiment analysis community is thriving. Unfortunately, not all submitted works could be represented in the workshop.

J.F. Sánchez-Rada, B. Schuller, G. Vulcu, C. A. Iglesias, P. Buitelaar, L. Devillers May 2016

[1] http://www.w3.org/community/sentiment/

[2] http://emotion-research.net/

[3] http://sspnet.eu/

Social Media

Monday 23 May, 9:10 – 10:30
Chair: J. Fernando Sánchez-Rada

Tweeting in the Debate about Catalan Elections

Cristina Bosco, Mirko Lai, Viviana Patti, Francisco M. Rangel Pardo and Paolo Rosso

The paper introduces a new annotated Spanish and Catalan data set for Sentiment Analysis about the Catalan separatism and the related debate held in social media at the end of 2015. It focuses on the collection of data, where we dealt with the exploitation in the debate of two languages, i.e. Spanish and Catalan, and on the design of the annotation scheme, previously applied in the development of other corpora about political debates, which extends a polarity label set by making available tags for irony and semantic oriented labels. The annotation process is presented and the detected disagreement discussed.

Emoji as Emotion Tags for Tweets

Ian D. Wood and Sebastian Ruder

In many natural language processing tasks, supervised machine learning approaches have proved most effective, and substantial effort has been made into collecting and annotating corpora for building such models. Emotion detection from text is no exception; however, research in this area is in its relative infancy, and few emotion annotated corpora exist to date. A further issue regarding the development of emotion annotated corpora is the difficulty of the annotation task and resulting inconsistencies in human annotations. One approach to address these problems is to use self-annotated data, using explicit indications of emotions included by the author of the data in question. We present a study of the use of unicode emoji as self-annotation of a Twitter user's emotional state. Emoji are found to be used far more extensively than hash tags and we argue that they present a more faithful representation of a user's emotional state. A substantial set of tweets containing emotion indicative emoji are collected and a sample annotated for emotions. The accuracy and utility of emoji as emotion labels are evaluated directly (with respect to annotations) and through trained statistical models. Results are cautiously optimistic and suggest further study of emotji usage.

Steam Review Dataset - new, large scale sentiment dataset

Antoni Sobkowicz and Wojciech Stokowiec

In this paper we present new binary sentiment classification dataset containing over 3,640,386 reviews from Steam User Reviews, with detailed analysis of dataset properties and initial results of sentiment analysis on collected data.

Corpora and Data Collection

Monday 23 May, 11:00 – 13:00
Chair: Viviana Patti

A Curated Corpus for Sentiment-Topic Analysis

Ebuka Ibeke, Chenghua Lin, Chris Coe, Adam Wyner, Dong Liu, Mohamad Hardyman Barawi and Noor Fazilla Abd Yusof

There has been a rapid growth of research interest in natural language processing that seeks to better understand sentiment or opinion expressed in text. However, most research focus on developing new models for opinion mining, with little efforts being devoted to the development of curated datasets for training and evaluation of these models. This work provides a manually annotated corpus of customer reviews, which has two unique characteristics. First, the corpus captures sentiment and topic information at both the review and sentence levels. Second, it is time-variant, which preserves the sentiment and topic dynamic information of the reviews. The annotation process was performed in a two-stage approach by three independent annotators, achieving a substantial level of inter-annotator agree- ments. In another set of experiments, we performed supervised sentiment classification using our manual annotations as gold-standard. Experimental results show that both Naive Bayes model and Support Vector Machine achieved more than 92

EmoCues-28: Extracting Words from Emotion Cues for a Fine-grained Emotion Lexicon

Jasy Liew Suet Yan and Howard R. Turtle

This paper presents a fine-grained emotion lexicon (EmoCues-28) consisting of words associated with 28 emotion categories. Words in the lexicon are extracted from emotion cues (i.e., any segment of text including words and phrases that constitute expression of an emotion) identified by annotators from a corpus of 15,553 tweets (microblog posts on Twitter). In order to distinguish between emotion categories at this fine-grained level, we introduce cue term weight and describe an approach to determine the primary and secondary terms associated with each emotion category. The primary and secondary terms form the foundation of our emotion lexicon. These terms can function as seed words to enrich the vocabulary of each emotion category. The primary terms can be used to retrieve synonyms or other semantically related words associated with each emotion category while secondary terms can be used capture contextual cues surrounding these terms.

A Bootstrapping Technique to Annotate Emotional Corpora Automatically

Lea Canales, Carlo Strapparava, Ester Boldrini and Patricio Martínez-Barco

In computational linguistics, the increasing interest of the detection of emotional and personality profiles has given birth to the creation of resources that allow the detection of these profiles. This is due to the large number of applications that the detection of emotion states can have, such as in e-learning environment or suicide prevention. The development of resources for emotional profiles can help to improve emotion detection techniques such as supervised machine learning, where the development of annotated corpora is crucial.

Generally, these annotated corpora are performed by a manual annotation process, a tedious and time-consuming task. Thus, research on developing automatic annotation processes has increased. Due to this, in this paper we propose a bootstrapping process to label an emotional corpus automatically, employing NRC Word-Emotion Association Lexicon (Emolex) to create the seed and generalised similarity measures to increase the initial seed. In the evaluation, the emotional model and the agreement between automatic and manual annotations are assessed.The results confirm the soundness of the proposed approach for automatic annotation and hence the possibility to create stable resources such as, an emotional corpus that can be employed on supervised machine learning for emotion detection systems.

A Multilingual Sentiment Corpus for Chinese, English and Japanese

Francis Bond, Tomoko Ohkuma, Luís Morgado da Costa, Yasuhide Miura, Rachel Chen, Takayuki Kuribayashi and Wenjie Wang

In this paper, we present the sentiment tagging of a multi-lingual corpus. The goal is to investigate how different languages encode sentiment, and compare the results with those given by existing resources. The results of annotating a corpus for both concept level and chunk level sentiment are analyzed.

Personality and User Modelling

Monday 23 May, 11:00 – 13:00
Chair: Björn Schuller

PACMAN: Psycho and Computational Framework of an Individual (Man)

Shivani Poddar, Sindhu Kiranmai Ernala and Navjyoti Singh

Several models have tried to understand the formation of an individual's distinctive character i.e. personality from the perspectives of multiple disciplines, including cognitive science, affective neuroscience and psychology. While these models (for eg. Big Five) have so far attempted to summarize the personality of an individual as a uniform, static image, no one model comprehensively captures the mechanisms which leads to the formation and evolution of personality traits over time. This mechanism of evolving personality is what we attempt to capture by means of our framework. Through this study, we leverage the Abhidhamma tradition of Buddhism to propose a theoretical model of an individual as a stochastic finite state machine. The machine models moment to moment states of consciousness of an individual in terms of a formal ontology of mental factors that constitute any individual. To achieve an empirical evaluation of our framework, we use social media data to model a user's personality as an evolution of his/her mental states (by conducting some psycho-linguistic inferences of their Facebook (FB) statuses). We further analyze the user's personality as a composition of these recurrent mental factors over a series of subsequent moments. As the first attempt to solve the problem of evolving personality explicitly, we also present a new dataset and machine learning module for analysis of mental states of a user from his/her social media data.

Telltale Trips: Personality Traits in Travel Blogs

Veronika Vincze1, Klára Hegedűs, Gábor Berend and Richárd Farkas

Here we present a corpus that contains blog texts about traveling. The main focus of our research is the personality trait of the person hence we do not just annotate opinions in the classical sense but we also mark those phrases that refer to the personality type of the author. We illustrate the annotation principles with several examples and we calculate inter-annotator agreement rates. In the long run, our main goal is to employ personality data in a real-world application, e.g. a recommendation system.

Linked Data and Semantics

Monday 23 May, 14:00 – 16:00
Chair: Ian D. Wood

Semantic Classification and Weight Matrices Derived from the Creation of Emotional Word Dictionary for Semantic Computing

Minsu Ko

This paper introduces a general creation method for an emotional word dictionary (EWD) which contains a semantic weight matrix (SWM) and a semantic classification matrix (SCM) which will be used as an efficient foundation for opinion mining. These two matrices are combined into a single n by 7 matrix called as a classification and weight matrix (CWM) in a machine-processable format. Such a matrix would also have applications in the field of semantic computing.This paper also details investigations which were performed in order to gather information on the efficiency of using CWM based on categorizing synonymous relations and frequencies. The multilingual extensibility of the EWD will benefit semantic processing of opinion mining as a generic linguistic resource which has an emotional ontology structure and linked data.

Towards a Common Linked Data Model for Sentiment and Emotion Analysis

J. Fernando Sánchez-Rada, Björn Schuller, Viviana Patti, Paul Buitelaar, Gabriela Vulcu, Felix Burkhardt, Chloé Clavel, Michael Petychakis and Carlos A. Iglesias

The different formats to encode information currently in use in sentiment analysis and opinion mining are heterogeneous and often custom tailored to each application. Besides a number of existing standards, there are additionally still plenty of open challenges, such as representing sentiment and emotion in web services, integration of different models of emotions or linking to other data sources. In this paper, we motivate the switch to a linked data approach in sentiment and emotion analysis that would overcome these and other current limitations. This paper includes a review of the existing approaches and their limitations, an introduction of the elements that would make this change possible, and a discussion of the challenges behind that change.

Beyond Text Analysis

Monday 23 May, 16:30 – 18:00
Chair: J. Fernando Sánchez-Rada

Empirical Mode Decomposition: A Data-Enrichment Perspective on Speech Emotion Recognition

Bin Dong, Zixing Zhang and Björn Schuller

To deal with the data scarcity problem for Speech Emotion Recognition, a novel data enrichment perspective is proposed in this paper by applying Empirical Mode Decomposition (EMD) on the existing labelled speech samples. In doing this, each speech sample is decomposed into a set of Intrinsic Mode Functions (IMFs) plus a residue by EMD. After that, we extract features from the primary IMFs of the speech sample. Each single classification model is trained first for the corresponding IMF. Then, all the trained models of the IMFs plus that of the original speech are combined together to classify the emotion by majority vote. Four popular emotional speech corpora and three feature sets are used in an extensive evaluation of the recognition performance of our proposed novel method. The results show that, our method can improve the classification accuracy of the prediction of valence and arousal with different significance levels, as compared to the baseline.

Automatic Detection of Textual Triggers of Reader Emotion in Short Stories

Rebekah Wegener, Christian Kohlschein, Sabina Jeschke and Björn Schuller

This position paper outlines our experimental design and platform development for remote data collection, annotation and analysis. The experimental design captures reader response to 5 short stories, including reading time, eye and gaze tracking, pupil dilation, facial gestures, combined physiological measures, spoken reflection, comprehension and reflection. Data will be gathered on a total corpus of 250 short stories and over 700 readers. We describe both the experiment design and the platform that will allow us to remotely crowd-source both reader response and expert annotation, as well as the means to analyse and query the resulting data. In the paper we outline our proposed approach for gaze-text linkage for remote low-quality webcam input and the proposed approach to the capture and analysis of low arousal affect data. The final platform will be open-source and fully accessible. We also plan to release all acquired data to the affective computing research community.

Domain Adaptation using Stock Market Prices to Refine Sentiment Dictionaries

Andrew Moore, Paul Rayson and Steven Young

As part of a larger project where we are examining the relationship and influence of news and social media on stock price, here we investigate the potential links between the sentiment of news articles about companies and stock price change of those companies. We describe a method to adapt sentiment word lists based on news articles about specific companies, in our case downloaded from the Guardian. Our novel approach here is to adapt word lists in sentiment classifiers for news articles based on the relevant stock price change of a company at the time of web publication of the articles. This adaptable word list approach is compared against the financial lexicon from Loughran and McDonald (2011) as well as the more general MPQA word list (Wilson et al., 2005). Our experiments investigate the need for domain specific word lists and demonstrate how general word lists miss indicators of sentiment by not creating or adapting lists that come directly from news about the company. The companies in our experiments are BP, Royal Dutch Shell and Volkswagen.

VisLR II: Visualization as Added Value in the Development, Use and Evaluation of Language Resources

23 May 2016

ABSTRACTS

Editors:

Annette Hautli-Janisz, Verena Lyding

Workshop Programme

09:00 – 10:35	Morning Session, Part I
09.00 – 09.05	Introduction
09.05 – 09.35	Paul Meurer, Victoria Rosén, Koenraad De Smedt *Interactive Visualizations in the INESS Treebanking Infrastructure*
09.35 – 10.05	Christin Schätzle, Dominik Sacha *Visualizing Language Change: Dative Subjects in Icelandic*
10.05 – 10.35	Annette Hautli-Janisz *See the Forest AND the Trees:* *Visual Verb Class Identification in Urdu/Hindi VerbNet*
10:35 – 11:00	*Coffee break*
11:00 – 12:40	Morning Session, Part II
11.00 – 11.30	Thomas Wielfaert, Kris Heylen, Dirk Speelman, Dirk Geeraerts *Visual Analytics for Distributional Semantic Model Comparisons*
11.30 – 12.00	Erik Tjong Kim Sang *Visualizing Literary Data*
12.00 – 12.30	Andrew Caines, Christian Bentz, Dimitrios Alikaniotis, Fridah Katushemererwe, Paula Buttery *The Glottolog Data Explorer: Mapping the World's Languages*
12:30 – 12:40	Closing

Workshop Organizers

Mennatallah El-Assady	University of Konstanz, Germany
Annette Hautli-Janisz	University of Konstanz, Germany
Verena Lyding	EURAC research Bozen/Bolzano, Italy

Workshop Programme Committee

Noah Bubenhofer	University of Zürich, Switzerland
Miriam Butt	University of Konstanz, Germany
Jason Chuang	Independent Researcher, USA
Christopher Collins	University of Ontario Institute of Technology, Canada
Chris Culy	Independent Consultant, USA
Gerhard Heyer	University of Leipzig, Germany
Kris Heylen	University of Leuven, Belgium
Daniel Keim	University of Konstanz, Germany
Steffen Koch	University of Stuttgart, Germany
Victoria Rosén	University of Bergen, Norway

Preface

This workshop aims at providing a follow-up forum to the successful first VisLR workshop at LREC 2014, which addresses visualization designers and users from computational and linguistic domains likewise. Since the last workshop, the concern with visualizing language data has further increased, as the recurrence of specialized symposia in the linguistic and NLP contexts show (cf. e.g. ACL workshop 2014, AVML 2014, Herrenhäuser Symposium 2014, QueryVis 2015). Moreover, the application of visualization techniques to various use cases is becoming ever more agile.

As a specialized subfield of information visualization, the visualization of language continues to face particular challenges: Language data is complex, only partly structured and, as with todays language resources, comes in large quantities. Moreover, due to the variety of data types, from textual data to spoken or signed language data, the challenges for visualization are necessarily varied. The overall challenge lies in breaking down the multidimensionality into intuitive visual features that enable an at-a-glance overview of the data. The second edition of the workshop therefore aims at advancing the field of linguistic visualization by particularly focusing on more advanced visualization techniques that represent the complexity of language and that contribute to resolving them.

Morning Session, Part I

Monday 23 May, 9:00 – 10:35
Chairperson: Verena Lyding

Interactive Visualizations in the INESS Treebanking Infrastructure

Paul Meurer, Victoria Rosén, Koenraad De Smedt

The visualization of syntactic analyses may be challenging due to the number of readings, the size and detail of the structures, and the interrelations between levels of linguistic description. We present a range of interactive visualization techniques applied to complex syntactic analyses in INESS, an online infrastructure for parsing and the annotation and exploration of syntactically annotated corpora (treebanks). Although INESS caters to many syntactic formalisms, we focus on LFG, which allows for multiple levels of syntactic structure, in particular c-structures and f-structures. Interactive dynamic renderings of the relations between components of these structures are presented, with options on the level of detail to be displayed. Furthermore, the disambiguation of sentences with multiple possible parses needs techniques for visualizing the differences between readings. For this purpose, we present and discuss packed representations, the interactive visualization of discriminants, and the previewing of disambiguation choices. The interactive querying of treebanks benefits from appropriate ways of displaying search results. We present the highlighting of matching items in matching sentences. We also present tabular overviews with frequencies of obtained variable values, as well as the inspection of matching structures without having to navigate away from the overview.

Visualizing Language Change: Dative Subjects in Icelandic

Christin Schätzle, Dominik Sacha

This paper presents a visualization tool for the analysis of diachronic multidimensional language data. Our tool was developed with respect to a corpus study of dative subjects in Icelandic based on the Icelandic Parsed Historical Corpus (Wallenberg et al., 2011) which investigates determining factors for the appearance of dative subjects in the history of Icelandic. The visualization provides an interactive access to the underlying multidimensional data and significantly facilitates the analysis of the complex diachronic interactions of factors at hand. We were able to identify various interactions of conditioning factors for dative subjects in Icelandic via the visualization tool and showed that dative subjects are increasingly associated with experiencer arguments in Icelandic across time. We also found that the rise of dative subjects with experiencer arguments is correlated with an increasing use of middle voice. This lexical semantic change argues against dative subjects as a Proto Indo-European inheritance. Moreover, the visualization helped us to draw conclusions about uncertainties and problems of our lexical semantic data annotation which will be revised for future work.

See the Forest AND the Trees: Visual Verb Class Identification in Urdu/Hindi VerbNet

Annette Hautli-Janisz

Constructing a lexical resource like VerbNet involves the crucial task of forming syntactically motivated subclasses within larger, semantically motivated verb classes. This is particularly challenging when the syntactic behavior of verbs varies considerably within a class, making well-motivated subclasses hard to establish by hand. The present paper shows that a visual clustering approach substantially facilitates the development process: Based on a careful theoretical analysis of the syntactic properties of Urdu/Hindi motion verbs, each verb can be represented by an 8-dimensional feature vector which serves as the basis for the automatic clustering with k-Means. In order to overcome the blackbox of machine learning approaches and to make the resulting clusters interpretable, the visualization reduces the high-dimensional verb vectors to two dimensions and visualizes the clusters and their members via an interactive interface, allowing for an inspection of the underlying data. This leads to the formation of subclasses in Urdu/Hindi VerbNet that are theoretically as well as computationally well-motivated.

Morning Session, Part II

Monday 23 May, 11:00 – 12:40
Chairperson: Annette Hautli-Janisz

Visual Analytics for Distributional Semantic Model Comparisons

Thomas Wielfaert, Kris Heylen, Dirk Speelman, Dirk Geeraerts

Distributional semantic models have shown to be a successful technique for Word Sense Disambiguation and Word Sense Induction tasks. However, these models, and more specifically the token-level variants, are extremely parameter-rich. We are still in the dark on how the different parameters can be efficiently set and even more on how to evaluate the outcome when no gold standard is readily available. To gain a better insight, we are developing a visual analytics approach which shows these models in two ways: a scatterplot matrix for inter-model parameter comparison and zoomable individual scatter plots allowing for more details on-demand. More specifically, we first use a scatterplot matrix to compare models with different parameter settings in a single view. This enables us to track selections of tokens over different models. On top of this, we create a scatter plot for each individual model, enriched with both model dependent and model independent features. This way, we can have a more in-depth visual analysis of what is going on and visualise the distinct properties or parameters of the individual model.

Visualizing Literary Data

Erik Tjong Kim Sang

We look at different aspects of Dutch magazines, both from the fields of literary studies and linguistic studies. We explore the background of authors with respect to birth locations, ages and gender, and also in how language use in the magazines evolved over a period of several decades. We have created several interactive visualizations which enable researchers to browse and analyze text data and their metadata. The design of these visualizations was nontrivial: invoking questions about how to deal with missing data and documents with multiple authors. The data required for some of the visualizations useful for researchers, were infeasible for the software architecture to generate within a reasonable time-span. In a case study, we look at some of the research questions that can be answered by the data visualizations and suggest another data view that could be interesting for literary research. Interesting topics for future research rely heavily on improvements of the search architecture used and including extra annotation layers to our text corpora.

The Glottolog Data Explorer: Mapping the World's Languages

Andrew Caines, Christian Bentz, Dimitrios Alikaniotis, Fridah Katushemererwe, Paula Buttery

We present THE GLOTTOLOG DATA EXPLORER, an interactive web application in which the world's languages are mapped using a JavaScript library in the 'Shiny' framework for R (Chang et al., 2016). The world's languages and major dialects are mapped using coordinates from the Glottolog database (Hammarström et al., 2016). The application is primarily intended to portray the endangerment status of the world's languages, and hence the default map shows the languages colour-coded for this factor. Subsequently, the user may opt to hide (or re-introduce) data subsets by endangerment status, and to resize the datapoints by speaker counts. Tooltips allow the user to view language family classification and links the user to the relevant Glottolog webpage for each entry. We provide a data table for exploration of the languages by various factors, and users may download subsets of the dataset via this table interface. The web application is freely available at http://cainesap.shinyapps.io/langmap

TA-COS 2016
Text Analytics for Cybersecurity and Online Safety

23 May 2016

ABSTRACTS

Editors:

Guy De Pauw, Ben Verhoeven, Bart Desmet, Els Lefever

Workshop Programme

14:00 - 15:00 – Keynote

Anna Vartapetiance and Lee Gillam, *Protecting the Vulnerable: Detection and Prevention of Online Grooming*

15:00 - 16:00 – Workshop Papers I

Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch and Derek Ruths, *A Web of Hate: Tackling Hateful Speech in Online Social Spaces*

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven and Walter Daelemans, *A Dictionary-based Approach to Racism Detection in Dutch Social Media*

16:00 - 16:30 – Coffee break

16:30 - 18:00 – Workshop Papers II

Adeola O Opesade, Mutawakilu A Tiamiyu and Tunde Adegbola, *Forensic Investigation of Linguistic Sources of Electronic Scam Mail: A Statistical Language Modelling Approach*

Richard Killam, Paul Cook, Natalia Stakhanova, *Android Malware Classification through Analysis of String Literals*

Shomir Wilson, Florian Schaub, Aswarth Dara, Sushain K. Cherivirala, Sebastian Zimmeck, Mads Schaarup Andersen, Pedro Giovanni Leon, Eduard Hovy and Norman Sadeh, *Demystifying Privacy Policies with Language Technologies: Progress and Challenges*

Workshop Organizers

Guy De Pauw	CLiPS - University of Antwerp, Belgium
Ben Verhoeven	CLiPS - University of Antwerp, Belgium
Bart Desmet	LT3 - Ghent University, Belgium
Els Lefever	LT3 - Ghent University, Belgium

Workshop Programme Committee

Walter Daelemans (chair)	CLiPS - University of Antwerp, Belgium
Veronique Hoste (chair)	LT3 - Ghent University, Belgium

Fabio Crestani	University of Lugano, Switserland
Maral Dadvar	Twente University, The Netherlands
Lee Gillam	University of Surrey, UK
Chris Emmery	University of Antwerp, Belgium
Giacomo Inches	Fincons Group AG, Switzerland
Eva Lievens	Ghent University, Belgium
Shervin Malmasi	Harvard Medical School, USA
Nick Pendar	Skytree Inc, USA
Karolien Poels	University of Antwerp, Belgium
Awais Rashid	Lancaster University, UK
Cynthia Van Hee	Ghent University, Belgium
Anna Vartapetiance	University of Surrey, UK

Preface

Text analytics technologies are being widely used as components in Big Data applications, allowing for the extraction of different types of information from large volumes of text. A growing number of research efforts is now investigating the applicability of these techniques for cybersecurity purposes. Many applications are using text analytics techniques to provide a safer online experience, by detecting unwanted content and behavior on the Internet. Other text analytics approaches attempt to detect illegal activity on online networks or monitor social media against the background of real-life threats. Alongside this quest, many ethical concerns arise, such as privacy issues and the potential abuse of such technology. The first workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS 2016) aims to bring together researchers that have an active interest in the development and application of such tools.

Following our call for papers, we received papers on a wide range of topics and with the help of our varied team of reviewers were able to select the most relevant and most interesting contributions. The first two papers that are presented at this workshop deal with the issue of identifying hate speech on social media. Saleem and colleagues describe a technique that automatically detects hateful communities, while Tulkens *et al.* present research on how to develop techniques that idenfity hateful words and phrases on social media.

In the second session of this workshop Opesada *et al.* present a study on the origin of 419 Scam e-mails, using text classification techniques that identify varieties of English. Killam *et al.* identify malware on the Android platform by using text analytics on the apps' binary files. Finally, Wilson *et al.* describe work on developing techniques that can aid people in understanding the often lengthy and complex terms of use that they agree to online.

We are very pleased with this wide variety of topics of the submitted papers and are furthermore very pleased to be able to kick off our workshop with a keynote lecture by Anna Vartapetiance of the University of Surrey's Centre for Cyber Security. She will present ongoing research on the automatic detection of online grooming. We are sure that the presentations at TA-COS 2016 will trigger fruitful discussions and will help foster the awareness of the increasingly important role text analytics can play in cybersecurity applications.

The TA-COS 2016 Organizers,
Guy De Pauw
Ben Verhoeven
Bart Desmet
Els Lefever
`www.ta-cos.org`

Keynote

23 May, 14:00 – 15:00
Chairperson: Guy De Pauw

Protecting the Vulnerable: Detection and Prevention of Online Grooming

Anna Vartapetiance and Lee Gillam

The 2012 EU Kids Online report revealed that 30% of 9-16 year-olds have made contact online with someone they did not know offline, and 9% have gone to an offline meeting with someone they first met online. The report suggests that this is “rarely harmful”, but is hoping against harm really the wisest course of action?

This talk presents details of our ongoing research and development on the prevention and detection of unsavoury activities which involve luring vulnerable people into ongoing abusive relationships. We will focus specifically on online grooming of children, discussing the potential to detect and prevent such grooming, and relevant theories and systems.

The talk will address some of the challenges involved with the practical implementation and use of such safeguards, in particular with respect to legal and ethical issues. We conclude by discussing the opportunities for protecting further groups vulnerable to grooming for emotional, financial, or other purposes.

Biography

Dr Anna Vartapetiance, is a graduate entrepreneur and postdoctoral researcher at the University of Surrey’s Centre for Cyber Security, as well as a committee member for the BCS ICT Ethics Specialist Group. Anna is currently on the advisory committee of the “Automatic Monitoring for Cyberspace Applications (AMiCA)” project as an international expert, and an associate of the Internet Service Providers Association (ISPA). She is also a member of the International Federation of Information Processing (IFIP) Special Interest Group 9.2.2 Framework on Ethics of Computing (SIG 9.2.2) and the Working Group on Social Accountability and Computing (WG 9.2).

Prior to her work as entrepreneur and postdoctoral researcher, she was awarded a PhD from the Department of Computer Science at the University of Surrey for her work on *deception detection* using Natural Language Processing (NLP) to develop (semi-)automated detection systems. Her research has found application in systems that enhance outcomes and issues related to national DNA databases, online gambling, virtual worlds and machine ethics and has been published in over 15 peer reviewed journal papers, proceedings and book chapters.

Currently, Anna is working on *parental controls* with a Child Online Safety project prototype for the detection / prevention of online grooming.

Workshop Papers I

23 May, 15:00 – 16:00
Chairperson: Véronique Hoste

A Web of Hate: Tackling Hateful Speech in Online Social Spaces

Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch and Derek Ruths

Online social platforms are beset with hateful speech - content that expresses hatred for a person or group of people. Such content can frighten, intimidate, or silence platform users, and some of it can inspire other users to commit violence. Despite widespread recognition of the problems posed by such content, reliable solutions even for detecting hateful speech are lacking. In the present work, we establish why keyword-based methods are insufficient for detection. We then propose an approach to detecting hateful speech that uses content produced by self-identifying hateful communities as training data. Our approach bypasses the expensive annotation process often required to train keyword systems and performs well across several established platforms, making substantial improvements over current state-of-the-art approaches.

A Dictionary-based Approach to Racism Detection in Dutch Social Media

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven and Walter Daelemans

We present a dictionary-based approach to racism detection in Dutch social media comments, which were retrieved from two public Belgian social media sites likely to attract racist reactions. These comments were labeled as racist or non-racist by multiple annotators. For our approach, three discourse dictionaries were created: first, we created a dictionary by retrieving possibly racist and more neutral terms from the training data, and then augmenting these with more general words to remove some bias. A second dictionary was created through automatic expansion using a `word2vec` model trained on a large corpus of general Dutch text. Finally, a third dictionary was created by manually filtering out incorrect expansions. We trained multiple Support Vector Machines, using the distribution of words over the different categories in the dictionaries as features. The best-performing model used the manually cleaned dictionary and obtained an F-score of 0.46 for the racist class on a test set consisting of unseen Dutch comments, retrieved from the same sites used for the training set. The automated expansion of the dictionary only slightly boosted the model's performance, and this increase in performance was not statistically significant. The fact that the coverage of the expanded dictionaries did increase indicates that the words that were automatically added did occur in the corpus, but were not able to meaningfully impact performance. The dictionaries, code, and the procedure for requesting the corpus are available at: `https://github.com/clips/hades`.

Workshop Papers II

23 May, 16:00 – 16:30
Chairperson: Walter Daelemans

Forensic Investigation of Linguistic Sources of Electronic Scam Mail: A Statistical Language Modelling Approach

Adeola O Opesade, Mutawakilu A Tiamiyu, Tunde Adegbola

Electronic handling of information is one of the defining technologies of the digital age. These same technologies have been exploited by unethical hands in what is now known as cybercrime. Cybercrime is of different types but of importance to the present study is the 419 Scam because it is generally (yet controversially) linked with a particular country - Nigeria. Previous research that attempted to unravel the controversy applied the Internet Protocol address tracing technique. The present study applied the statistical language modelling technique to investigate the propensity of Nigeria's involvement in authoring these fraudulent mails. Using a hierarchical modelling approach proposed in the study, 28.85% of anonymous electronic scam mails were classified as being from Nigeria among four other countries. The study concluded that linguistic cues have potentials of being used for investigating transnational digital breaches and that electronic scam mail problem cannot be pinned down to Nigeria as believed generally, though Nigeria could be one of the countries that are prominent in authoring such mails.

Android Malware Classification through Analysis of String Literals

Richard Killam, Paul Cook and Natalia Stakhanova

As the popularity of the Android platform grows, the number of malicious apps targeting this platform grows along with it. Accordingly, as the number of malicious apps increases, so too does the need for an automated system which can effectively detect and classify these apps and their families. This paper presents a new system for classifying malware by leveraging the text strings present in an app's binary files. This approach was tested using over 5,000 apps from 14 different malware families and was able to classify samples with over 99% accuracy while maintaining a false positive rate of 2.0%.

Demystifying Privacy Policies with Language Technologies: Progress and Challenges

Shomir Wilson, Florian Schaub, Aswarth Dara, Sushain K. Cherivirala, Sebastian Zimmeck, Mads Schaarup Andersen, Pedro Giovanni Leon, Eduard Hovy and Norman Sadeh

Privacy policies written in natural language are the predominant method that operators of websites and online services use to communicate privacy practices to their users. However, these documents are infrequently read by Internet users, due in part to the length and complexity of the text. These factors also inhibit the efforts of regulators to assess privacy practices or to enforce standards. One proposed approach to improving the status quo is to use a combination of methods from crowdsourcing, natural language processing, and machine learning to extract details from privacy policies and present them in an understandable fashion. We sketch out this vision and describe our ongoing work to bring it to fruition. Further, we discuss challenges associated with bridging the gap between the contents of privacy policy text and website users' abilities to understand those policies. These challenges are motivated by the rich interconnectedness of the problems as well as the broader impact of helping Internet users understand their privacy choices. They could also provide a basis for competitions that use the annotated corpus introduced in this paper.

LDL 2016
5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources

24 May 2016

ABSTRACTS

Editors:

John P. McCrae, Christian Chiarcos, Elena Montiel Ponsoda, Thierry Declerck, Petya Osenova, Sebastian Hellmann

Workshop Programme

09:00 – 09:30 Welcome and introduction

09:30 – 10:30 Invited talk by Damir Cavar (Indiana University and The LINGUIST List), *On the role of Linked Open Language Data for Language Documentation, Linguistic Research, and Speech and Language Technologies*

10:30 – 11:00 Coffee break

11:00 – 11:30 Vladimir Alexiev and Gerard Casamayor, *FN goes NIF: Integrating FrameNet in the NLP Interchange Format*
11:30 – 12:00 Frank Abromeit, Christian Chiarcos, Christian Fäth and Maxim Ionov, *Linking the Tower of Babel: Modelling a Massive Set of Etymological Dictionaries as RDF*
12:00 – 12:20 Jim O Regan, Kevin Scannell and Elaine Uí Dhonnchadha, *lemonGAWN: WordNet Gaeilge as Linked Data*
12:20 – 12:40 Vít Baisa, Sara Može and Irene Renau, *Linking Verb Pattern Dictionaries of English and Spanish*

12:40 – 14:00 Lunch break

14:00 – 14:40 Poster Session

Fahad Khan, Javier Díaz-Vera and Monica Monachini, *The Representation of an Old English Emotion Lexicon as Linked Open Data*
Elena Gonzalez-Blanco, Gimena Del Rio Riande and Clara Martínez Cantón, *Linked open data to represent multilingual poetry collections. A proposal to solve interoperability issues between poetic repertoires*
Sotiris Karampatakis, Sofia Karampataki, Charalampos Bratsas and Ioannis Antoniou, *Linked Open Lexical Resources for the Greek Language*
Invited posters from related projects and initiatives

14:40 – 15:10 Petya Osenova and Kiril Simov, *Linked Open Data Dataset from Related Documents*
15:10 – 15:40 Andrea Salfinger, Caroline Salfinger, Birgit Pröll, Werner Retschitzegger and Wieland Schwinger, *Pinpointing the Eye of the Hurricane - Creating a Gold-Standard Corpus for Situative Geo-Coding of Crisis Tweets Based on Linked Open Data*
15:40 – 16:00 Thierry Declerck, *Representation of Polarity Information of Elements of German Compound Words*

16:00 – 16:30 Coffee break

16:30 – 16:50 Vanya Dimitrova, Christian Fäth, Christian Chiarcos, Heike Renner-Westermann and Frank Abromeit, *Building an ontological model of the BLL Thesaurus: First step towards an interface with the LLOD cloud*
16:50 – 17:10 Claus Zinn and Thorsten Trippel, *Enhancing the Quality of Metadata by using Authority Control*
17:10 – 17:30 Christian Chiarcos, Christian Fäth and Maria Sukhareva, *Developing and Using Ontologies of Linguistic Annotation*

17:30 – 18:00 Wrap up

Workshop Organizers

John P. McCrae	National University of Ireland, Galway, Ireland
Christian Chiarcos	Goethe University Frankfurt, German
Elena Montiel Ponsoda	Universidad Politécnica de Madrid, Spain
Thierry Declerck	Saarland University, Germany
Petya Osenova	IICT-BAS, Bulgaria
Sebastian Hellmann	AKSW/KILT, Universität Leipzig, Germany
Julia Bosque-Gil	Universidad Politécnica de Madrid, Spain
Bettina Klimek	AKSW/KILT, Universität Leipzig, Germany

Workshop Programme Committee

Guadalupe Aguado	Universidad Politécnica de Madrid, Spain
Núria Bel	Universitat Pompeu Fabra, Spain
Claire Bonial	University of Colorado at Boulder, USA
Paul Buitelaar	National University of Ireland, Galway, Ireland
Steve Cassidy	Macquarie University, Australia
Nicoletta Calzolari	ILC-CNR, Italy
Damir Cavar	Eastern Michigan University, USA
Philipp Cimiano	Bielefeld University, Germany
Gerard de Melo	Tsinghua University, China
Alexis Dimitriadis	Universiteit Utrecht, The Netherlands
Judith Eckle-Kohler	Technische Universität Darmstadt, Germany
Francesca Frontini	ILC-CNR, Italy
Jeff Good	University at Buffalo, USA
Asunción Gómez Pérez	Universidad Politécnica de Madrid, Spain
Jorge Gracia	Universidad Politécnica de Madrid, Spain
Yoshihiko Hayashi	Waseda University, Japan
Nancy Ide	Vassar College, USA
Fahad Khan	ILC-CNR, Italy
Vanessa Lopez	IBM Europe, Ireland
Steven Moran	Universität Zürich, Switzerland/ Ludwig MaximilianUniversity, Germany
Roberto Navigli	University of Rome, “La Sapienza”, Italy
Sebastian Nordhof	Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
Antonio Pareja-Lora	Universidad Complutense Madrid, Spain
Maciej Piasecki	Wroclaw University of Technology, Poland
Francesca Quattri	Hong Kong Polytechnic University, Hong Kong
Mariano Rico	Universidad Politécnica de Madrid, Spain
Laurent Romary	INRIA, France
Felix Sasaki	Deutsches Forschungszentrum für Künstliche Intelligenz, Germany
Andrea Schalley	Griffith University, Australia
Gilles Sérraset	Joseph Fourier University, France
Kiril Simov	Bulgarian Academy of Sciences, Sofia, Bulgaria

Milena Slavcheva	JRC-Brussels, Belgium
Aitor Soroa	University of the Basque Country, Spain
Armando Stellato	University of Rome, Tor Vergata, Italy
Cristina Vertan	University of Hamburg, Germany
Piek Vossen	Vrije Universiteit Amsterdam, The Netherlands

Preface

Since its establishment in 2012, the Linked Data in Linguistics (LDL) workshop series has become the major forum for presenting, discussing and disseminating technologies, vocabularies, resources and experiences regarding the application of **Semantic Web standards and the Linked Open Data paradigm to language resources** in order to facilitate their visibility, accessibility, interoperability, reusability, enrichment, combined evaluation and integration. The Linked Data in Linguistics workshop series is organized by the Open Linguistics Working Group of the Open Knowledge Foundation, and has contributed greatly to the development of the Linguistic Linked Open Data (LLOD) cloud. This workshop builds on the existing success of previous instances of this workshop over the last four years, firstly at the 34th Annual Conference of the German Linguistics Society (DGfS) in 2012, followed by a second appearance at the 6th Annual Conference on Generative Approaches to the Lexicon (GLCON). In 2014, the workshop was held at the previous edition of LREC in Reykjavik, where we attracted a very large number of interested participants. Last year, the workshop was co-located with ACL-IJCNLP 2015 in Beijing, China.

Publishing language resources under open licenses and linking them together has been an area of increasing interest in academic circles, including applied linguistics, lexicography, natural language processing and information technology, and to facilitate exchange of knowledge and information across boundaries between disciplines as well as between academia and the IT business. By collocating the 5th edition of the workshop series with LREC, we encourage this interdisciplinary community **to present and to discuss use cases, experiences, best practices, recommendations and technologies** among each other and in interaction with the language resource community. We particularly invite contributions discussing the application of the Linked Open Data paradigm to linguistic data as it might provide an important step towards making linguistic data: i) easily and uniformly **queryable**, ii) **interoperable** and iii) **sharable** over the Web using open standards such as the HTTP protocol and the RDF data model. While it has been shown that linked data has significant value for the management of language resources in the Web, the practice is still far from being an accepted standard in the community. Thus it is important that we continue to push the development and adoption of linked data technologies among creators of language resources. In particular linked data's ability to increase the **quality**, **interoperability** and **availability** of data on the Web has led to us focus on **managing, improving and using language resources on the Web** as a key focus for this year's workshop.

Session 1: Invited talk

Tuesday 24 May, 09:30 – 10:30
Chairperson: Christian Chiarcos

On the role of Linked Open Language Data for Language Documentation, Linguistic Research, and Speech and Language Technologies

Damir Cavar

We will discuss the potential roles of Linked and Open Language Data in the fields of documentary linguistics, theoretical linguistic research, and speech and language technology engineering, its potential in establishing new synergies between the fields, and enabling emerging new research questions and results. We will address questions on interoperability and standards for resource annotation, linking, and sharing of open language resources, and how this can improve accessibility and common restrictions on the half-life of language resources in particular for digital language archives. The questions that we want to address are, among others: Can a linked open language data approach shift digital language archives from a situation of being *language graveyards* to a useful resource for various sub-field of linguistics and speech and language related technologies? How can linking and openness be achieved given the various opposing constraints and needs with respect to language data and resources?

Session 2: oral presentations

Tuesday 24 May, 11:00 – 12:40
Chairperson: Elena Montiel Ponsoda

FN goes NIF: Integrating FrameNet in the NLP Interchange Format

Vladimir Alexiev and Gerard Casamayor

FrameNet (FN) is a large-scale lexical database for English developed at ICSI Berkeley that describes word senses in terms of Frame semantics. FN has been converted to RDF LOD by ISTC-CNR, together with a large corpus of text annotated with FN. NIF is an RDF/OWL format and protocol for exchanging text annotations between NLP tools as Linguistic Linked Data. This paper reviews the FN-LOD representation, compares it to NIF, and describes a simple way to integrate FN in NIF, which does not use any custom classes or properties.

Linking the Tower of Babel: Modelling a Massive Set of Etymological Dictionaries as RDF

Frank Abromeit, Christian Chiarcos, Christian Fäth and Maxim Ionov

This paper describes the development of a Linked Data representation of the Tower of Babel (Starling), a major resource for short- and long-range etymological relations. Etymological dictionaries are highly multilingual by design, they usually involve cross-references to additional monolingual and etymological dictionaries, and thus represent an ideal application of Linked Data principles. So far, however, the Linguistic Linked Open Data (LLOD) community rarely addressed etymological relations. In line with state-of-the-art LLOD practice, we represent Starling data in accordance with the lemon vocabulary developed by the W3C Ontolex Community Group, we discuss the state of the art, experiences and suggest extensions for etymological dictionaries.

WordNet Gaeilge as Linked Data

Jim O Regan, Kevin Scannell and Elaine Uí Dhonnchadha

We introduce lemonGAWN, a conversion of WordNet Gaeilge, a wordnet for the Irish language, with synset relations projected from EuroWordNet. lemonGAWN is linked to the English WordNet, as well as the wordnets for four Iberian languages covered by MCR, and additionally contains links to both the Irish and English editions of DBpedia.

Linking Verb Pattern Dictionaries of English and Spanish

Vít Baisa, Sara Može and Irene Renau

The paper presents the first step in the creation of a new multilingual and corpus-driven lexical resource by means of linking existing monolingual pattern dictionaries of English and Spanish verbs. The two dictionaries were compiled through Corpus Pattern Analysis (CPA) – an empirical procedure in corpus linguistics that associates word meaning with word use by means of analysis of phraseological patterns and collocations found in corpus data. This paper provides a first look into a number of practical issues arising from the task of linking corresponding patterns across languages via both manual and automatic procedures. In order to facilitate manual pattern linking, we implemented a heuristic-based algorithm to generate automatic suggestions for candidate verb pattern pairs, which obtained 80% precision. Our goal is to kick-start the development of a new resource for verbs that can be used by language learners, translators, editors and the research community alike.

Session 3: poster presentations

24 May 2016, 14:00 – 14:40
Chairperson: Thierry Declerck

The Representation of an Old English Emotion Lexicon as Linked Open Data

Fahad Khan, Javier Díaz-Vera and Monica Monachini

We present the ongoing conversion of a lexicon of emotion terms in Old English (OE) into RDF using an extension of lemon called lemonDIA and which we briefly describe. We focus on the translation of the subset of the lexicon dealing with terms for shame and guilt and give a number of illustrative examples.

Linked Open Data to represent multilingual Poetry Collections. A Proposal to solve Interoperability Issues between poetic Repertoires

Elena Gonzalez-Blanco, Gimena Del Rio Riande and Clara Martínez Cantón

This paper describes the creation of a poetic ontology in order to use it as a basis to link the different databases and projects working on metrics and poetry. It is built on the model of the Spanish digital repertoire ReMetCa, but its aim is to be enlarged and improved in order to fit under every poetic system. The conceptual semantic model, written in OWL, includes classes and metadata from standard ontological models related to humanities fields (such as CIDOC or Dublin Core), and adds specific elements and properties to describe poetic phenomena. Its final objective is to interconnect, reuse and locate data disseminated through poetic repertoires, in order to boost interoperability among them.

Linked Open Lexical Resources for the Greek Language

Sotiris Karampatakis, Sofia Karampataki, Charalampos Bratsas and Ioannis Antoniou

The continuous rise of information technology has led to a remarkable quantity of linguistic data that are accessible on the Web. Linguistic resources are more useful when linked but their distribution on the Web in various or closed formats makes it difficult to interlink with one another. So, they often end up restricted in data "silos". To overcome this problem Linked Data offers a way of publishing data using Web technologies in order to make feasible the connection of data coming from different sources. This paper presents how Greek WordNet was converted and published using a Resource Description Framework (RDF) model in order to expose as Linked Open Data.

Session 4: oral presentations

Tuesday 24 May, 14:40 – 16:00
Chairperson: John P. McCrae

Linked Open Data Dataset from Related Documents

Petya Osenova and Kiril Simov

In the paper we present a methodology for the creation of a LOD dataset over domain documents. The domain is European Law in its connection to national laws. The documents are interlinked over the web. They are linked also to other Linked Open Data datasets (such as GeoNames). The data includes five languages: Bulgarian, English, French, Italian and German. Thus, the paper discusses the first step towards the creation of a domain corpus with linguistically linked documents, namely - the reference linking among the documents and their linguistic processing.

Pinpointing the Eye of the Hurricane - Creating a Gold-Standard Corpus for Situative Geo-Coding of Crisis Tweets Based on Linked Open Data

*Andrea Salfinger, Caroline Salfinger, Birgit Pröll, Werner Retschitzegger and Wieland Schwinge*r

Crisis management systems would benefit from exploiting human observations of disaster sites shared in near-real time via microblogs, however, utterly require location information in order to make use of these. Whereas the popularity of microblogging services, such as Twitter, is on the rise, the percentage of GPS-stamped Twitter microblog articles (i.e., tweets) is stagnating. Geo-coding techniques, which extract location information from text, represent a promising means to overcome this limitation. However, whereas geo-coding of news articles represents a well-studied area, the brevity, informal nature and lack of context encountered in tweets introduces novel challenges on their geo-coding. Few efforts so far have been devoted to analyzing the different types of geographical information users mention in tweets, and the challenges of geo-coding these in the light of omitted context by exploiting situative information. To overcome this limitation, we propose a gold-standard corpus building approach for evaluating such situative geo-coding, and contribute a human-curated, geo-referenced tweet corpus covering a real-world crisis event, suited for benchmarking of geo-coding tools. We demonstrate how incorporating a semantically rich Linked Open Data resource facilitates the analysis of types and prevalence of geospatial information encountered in crisis-related tweets, thereby highlighting directions for further research.

Representation of Polarity Information of Elements of German Compound Words

Thierry Declerck

We present on-going work on using formal representation frameworks for encoding polarity information that can be attached to elements of German compound words. As a departure point we have a polarity lexicon for German words that was compiled and ranked on the basis of the integration of four pre-existing polarity lexicons that were available in different formats. As for the formal representation frameworks we are considering for the encoding of the lexical data the lexicon model for ontologies (lemon), more specifically its modules ontolex (Ontology-lexicon interface) and decomp (Decomposition), which have been developed in the context of the W3C Ontology-Lexica Community Group. For the encoding of the polarity information we adopt a slightly modified version of the Marl ontological modelling, developed at the Universidad Politécnica de Madrid.

Session 5: oral presentations

Tuesday 24 May, 16:30 – 17:30
Chairperson: Bettina Klimek

Building an ontological model of the BLL Thesaurus: First step towards an interface with the LLOD cloud Documents

Vanya Dimitrova, Christian Fäth, Christian Chiarcos, Heike Renner-Westermann and Frank Abromeit

This paper describes the on-going efforts to position the Bibliography of Linguistic Literature (BLL) within the wider context of Linguistic Linked Open Data (LLOD) and to enhance the functionality of the Lin|gu|is|tik portal, a virtual library for the field of linguistics, with an LOD interface. Being the connecting point between the portal and LLOD cloud, the BLL Thesaurus has to fulfil certain formal and conceptual requirements, i.e., it has to be remodelled as ontology. The remodelling of the BLL Thesaurus is the main subject of the paper. Starting with the specificity of the Thesaurus, its scope and nature, we describe our general methodological approach and design solutions. We present the basic ontological framework and give concrete examples from our work in progress. Challenging cases from the domains of morphology and syntax depict the complexity of the task. Additionally, we specify the next steps towards an LOD interface and long term perspectives.

Enhancing the Quality of Metadata by using Authority Control

Claus Zinn and Thorsten Trippel

The Component MetaData Infrastructure (CMDI) is the dominant framework for describing language resources according to ISO 24622. Within the CLARIN world, CMDI has become a huge success. The Virtual Language Observatory (VLO) now holds over 800.000 resources, all described with CMDI-based metadata. With the metadata being harvested from about thirty centres, there is a considerable amount of heterogeneity in the data. In part, there is some use of controlled vocabularies to keep data heterogeneity in check, say when describing the type of a resource, or the country the resource is originating from. However, when CMDI data refers to the names of persons or organisations, strings are used in a rather uncontrolled manner. Here, the CMDI community can learn from libraries and archives who maintain standardised lists for all kinds of names. In this paper, we advocate the use of freely available authority files that support the unique identification of persons, organisations, and more. The systematic use of authority records enhances the quality of the metadata, hence improves the faceted browsing experience in the VLO, and also prepares the sharing of CMDI-based metadata with the data in library catalogues.

Developing and Using Ontologies of Linguistic Annotation

Christian Chiarcos, Christian Fäth and Maria Sukhareva

This paper describes the Ontologies of Linguistic Annotation (OLiA) as one of the data sets currently available as part of Linguistic Linked Open Data (LLOD) cloud. Within the LLOD cloud, the OLiA ontologies serve as a reference hub for annotation terminology for linguistic phenomena on a great band-width of languages, they have been used to facilitate interoperability and information integration of linguistic annotations in corpora, NLP pipelines, and lexical-semantic resources and mediate their linking with multiple community-maintained terminology repositories.

The purpose of this paper is to provide an overview over progress, recent applications and prospective developments, and to introduce two novel applications of OLiA.

GLOBALEX 2016
Lexicographic Resources for Human Language Technology

24 May 2016

ABSTRACTS

Editors:

Ilan Kernerman, Iztok Kosem, Simon Krek, Lars Trap-Jensen

Workshop Programme

09:00 – 09:10 Introduction by Ilan Kernerman and Simon Krek

09:10 – 10:30 Session 1
Patrick Hanks, *A common-sense paradigm for linguistic research*
Pamela Faber, Pilar León-Araúz and Arianne Reimerink, *EcoLexicon: New features and challenges*
Sara Carvalho, Rute Costa and Christophe Roche, *Ontoterminology meets lexicography: The Multimodal Online Dictionary of Endometriosis (MODE)*
Gregory Grefenstette and Lawrence Muchemi, *Determining the characteristic vocabulary for a specialized dictionary using Word2vec and a directed crawler*

10:30 – 11:00 Coffee break

11:00 – 12:40 Session 2
Malin Ahlberg, Lars Borin, Markus Forsberg, Olof Olsson, Anne Schumacher and Jonatan Uppström, *Karp: Språkbanken's open lexical infrastructure*
Ivelina Stoyanova, Svetla Koeva, Maria Todorova and Svetlozara Leseva, *Semi-automatic compilation of a very large multiword expression dictionary for Bulgarian*
Raffaele Simone and Valentina Piunno, *CombiNet: Italian word combinations in an online lexicographic tool*
Irena Srdanovic and Iztok Kosem, *GDEX for Japanese: Automatic extraction of good dictionary example candidates*
Jana Klimová, Veroniká Kolářová and Anna Vernerová, *Towards a corpus-based valency of Czech nouns*

12:40 – 14:20 Lunch break

14:20 – 16:00 Session 3
Jan Hajic, Eva Fucikova, Jana Sindlerova and Zdenka Uresova, *Verb argument pairing in a Czech-English parallel treebank*
Sonja Bosch and Laurette Pretorius, *The role of computational Zulu verb morphology in multilingual lexicographic applications*
Martin Benjamin, *Toward a global online living dictionary: A model and obstacles for big linguistic data*
Luis Morgado da Costa, Francis Bond and František Kratochvil, *Linking and disambiguating Swadesh texts*
Luis Espinoza Anke, Roberto Carlini, Horacio Saggion and Francesco Ronzano, *DEFEXT: A semi-supervised definition extraction tool*

16:00 – 16:30 Coffee break

16:30 – 17:10 Session 4
Julia Bosque-Gil, Jorge Gracia, Elena Montiel-Ponsoda and Guadalupe Aguado-de-Cea, *Modelling multilingual lexicographic resources for the web of data: The K Dictionaries case*
Ivett Benyeda, Péter Koczka and Tamás Váradi, *Creating seed lexicons for under-resourced languages*

17:10 – 18:00 GLOBALEX Discussion

Workshop Organizers

Andrea Abel	EURALEX
Ilan Kernerman*	ASIALEX
Steven Kleinedler	DSNA
Iztok Kosem	eLex
Simon Krek*	eLex
Julia Miller	AUSTRALEX
Maropeng Victor Mojela	AFRILEX
Danie J. Prinsloo	AFRILEX
Rachel Edita O. Roxas	ASIALEX
Lars Trap-Jensen	EURALEX
Luanne von Schneidemesser	DSNA
Michael Walsh	AUSTRALEX

* Co-chairs of the Organising Committee

Workshop Programme Committee

Michael Adams	Indiana University
Philipp Cimiano	University of Bielefeld
Janet DeCesaris	Universitat Pompeu Fabra
Thierry Declerck	German Research Center for Artificial Intelligence
Anne Dykstra	Fryske Akademie
Edward Finegan	University of Southern California
Thierry Fontenelle	Translation Center for the Bodies of the EU
Polona Gantar	University of Ljubljana
Alexander Geyken	Berlin-Brandenburg Academy of Sciences and Humanities
Rufus Gouws	Stellenbosch University
Jorge Gracia	Madrid Polytechnic University
Orin Hargraves	University of Colorado
Ulrich Heid	Hildesheim University
Chu-Ren Huang	Hong Kong Polytechnic University
Miloš Jakubíček	Lexical Computing – Sketch Engine
Jelena Kallas	Institute of Estonian Language
Ilan Kernerman	K Dictionaries
Annette Klosa	German Language Institute
Iztok Kosem	Trojina
Simon Krek	“Jožef Stefan” Institute
Robert Lew	Adam Mickiewicz University
Marie Claude l’Homme	Université de Montréal
Nikola Ljubešić	University of Zagreb
Stella Markantonatou	Institute for Language and Speech Processing ATHENA
John McCrae	National University of Ireland Galway
Roberto Navigli	Sapienza University of Rome
Vincent Ooi	National University of Singapore
Michael Rundell	Lexicography Masterclass

Mary Salisbury	Massey University
Adam Smith	Macquarie University
Pius ten Hacken	Innsbruck University
Carole Tiberius	Institute of Dutch Lexicology
Yukio Tono	Tokyo University of Foreign Studies
Lars Trap-Jensen	Society for Danish Language and Literature
Tamás Váradi	Hungarian Academy of Sciences
Elena Volodina	Gothenburg University
Eveline Wandl-Vogt	Austrian Academy of Sciences
Shigeru Yamada	Waseda University

Introduction

The field of lexicography has been shifting to digital media, with effect on all stages of research, development, design, evaluation, publication, marketing and usage. Modern lexicographic content is created with help of dictionary writing tools, corpus query systems and QA applications, and becomes more easily accessible and useful for integration with numerous LT solutions, as part of bigger knowledge systems and collaborative intelligence.

At the same time, extensive interlinked language resources, primarily intended for use in Human Language Technology (HLT), are being created through projects, movements and initiatives, such as Linguistic Linked (Open) Data (LLOD), meeting requirements for optimal use in HLT, e.g. unique identification and use of web standards (RDF or JSON-LD), leading to better federation, interoperability and flexible representation. In this context, lexicography constitutes a natural and vital part of the LLOD scheme, currently represented by wordnets, FrameNets, and HLT-oriented lexicons, ontologies and lexical databases. However, a new research paradigm and common standards are still lacking, and so are common standards for the interoperability of lexicography with HLT applications and systems.

The aim of this workshop is to explore the development of global standards for the evaluation of lexicographic resources and their incorporation with new language technology services and other devices. The workshop is the first-ever joint initiative by all the major continental lexicography associations, seeking to promote cooperation with related fields of HLT for all languages worldwide, and it is intended to bridge various existing gaps within and among such different research fields and interest groups. The target audience includes lexicographers, computational and corpus linguists and researchers working in the fields of HLT, Linked Data, the Semantic Web, Artificial Intelligence, etc.

GLOBALEX 2016 is sponsored by the five existing continental lexicography associations and the international conferences on electronic lexicography:

- AFRILEX – The African Association for Lexicography
- ASIALEX – The Asian Association for Lexicography
- AUSTRALEX – The Australasian Association for Lexicography
- DSNA – The Dictionary Society of North America
- EURALEX – The European Association for Lexicography
- eLex – Electronic Lexicography in the 21st Century conferences

This workshop constitutes the initial step in forming GLOBALEX – a global constellation for all continental, regional, local, topical or special interest communities concerned with lexicography.

GLOBALEX will promote knowledge sharing and cooperation among its members and with other parties concerned with language and linguistics. It will aim to establish global standards for the creation, evaluation, dissemination and usage of lexicographic resources and solutions, and for the interoperability of lexicography with other relevant disciplines and branches of the HLT academe and industry worldwide.

Tuesday 24 May, 9:10 – 9:30
Chairperson: Michael Rundell

A common-sense paradigm for linguistic research

Patrick Hanks

During the past century, studies in linguistics have contributed greatly to the understanding of syntax and phonology, while philosophers have illuminated functions of language. This paper proposes that the time has now come to focus on something more obvious—the empirical investigation of meaning by studying word use and phraseology. It seems unlikely that even the most abstract theoretician would deny that a natural language consists primarily of words (among other phenomena), and that the primary function of word use is to make meaningful statements. But word use has not yet been adequately investigated. Astonishingly, there is still much to be discovered about the mechanisms of meaningful word use. This has been a topic in corpus linguistics since its inception a quarter of a century ago, but a firmer theoretical foundation is needed. When we attempt to build one, some surprising issues rapidly become apparent. Corpus linguistics has established, in the words of John Sinclair, that "Many if not most meanings require the presence of more than one word for their normal realization. [...] Patterns of co-selection among words, which are much stronger than any description has yet allowed for, have a direct connection with meaning." This limpid observation implies a vast programme of research in collocation and phraseology in all the world's languages. How exactly do people put words together in order to create meanings? For many years, I (a Sinclairian) have claimed, "Words don't have meaning; they only have meaning potential." How can such a claim be reconciled with the common-sense observation, "Of course words have meaning: I know what the word elephant means, because that is part of my competence as a speaker of English"? A first answer would be to to cite phrases such as the elephant in the room and a white elephant. You will say that these are idioms, which are different from the word in isolation. I will then ask, what about phraseology like "The competition has come to resemble an elephant's graveyard" and "[His] bitter mien suggests inner torment and an elephant's memory for grievance." It seems that there is a lot more going on than mere concatenation. Such phrases exploit attributed properties (real or imagined) of the noun. We might agree that concrete nouns such as elephant have a central default meaning, but in phrase after phrase corpus evidence show that something more dramatic is going on, not adequately accounted for in current dictionaries. Some other questions that arise (and that can be investigated empirically through corpus analysis) are: • Are meanings evanescent? Are they events that pass away as soon as uttered, or are they more stable abstract entities? • Are noun meanings different in kind from verb or adjective meanings? • How important is context, and what counts are relevant context? • Does phraseology determine meaning? • What is the nature of linguistic creativity? • What is the nature of linguistic salience? Some senses of a word are clearly more frequent than others, but should we distinguish between social salience (frequency) and cognitive salience (recallability)? • Is scientific discourse, with its heavy reliance on technical terminology, different in kind from ordinary language? An adequate research paradigm for investigating meaning must provide a framework for investigating such questions. This paper will suggest how such questions might be approached.

Tuesday 24 May, 9:30 – 9:50
Chairperson: Michael Rundell

EcoLexicon: New features and challenges

Pamela Faber, Pilar León-Araúz and Arianne Reimerink

EcoLexicon is a terminological knowledge base on the environment with terms in six languages: English, French, German, Greek, Russian, and Spanish. It is the practical application of Frame-based Terminology, which uses a modified version of Fillmore's Frames coupled with premises from Cognitive Linguistics to configure specialized domains on the basis of definitional templates and create situated representations for specialized knowledge concepts. The specification of the conceptual structure of (sub)events and the description of the lexical units are the result of a top-down and bottom-up approach that extracts information from a wide range of resources. This includes the use of corpora, the factorization of definitions from specialized resources and the extraction of conceptual relations with knowledge patterns. Similarly to a specialized visual thesaurus, EcoLexicon provides entries in the form of semantic networks that specify relations between environmental concepts. All entries are linked to a corresponding (sub)event and conceptual category. In other words, the structure of the conceptual, graphical, and linguistic information relative to entries is based on an underlying conceptual frame. Graphical information includes photos, images, and videos, whereas linguistic information not only specifies the grammatical category of each term, but also phraseological, and contextual information. The TKB also provides access to the specialized corpus created for its development and a search engine to query it. One of the challenges of EcoLexicon in the near future is its inclusion in the Linguistic Linked Open Data Cloud.

Tuesday 24 May, 9:50 – 10:10
Chairperson: Michael Rundell

Ontoterminology meets lexicography: The Multimodal Online Dictionary of Endometriosis (MODE)

Sara Carvalho, Rute Costa and Christophe Roche

With the advent of the Semantic Web and, more recently, of the Linked Data initiative, the need to operationalise lexicographic resources, i.e. to represent them in a computer-readable format, has become increasingly important, as it contributes to pave the way to the ultimate goal of interoperability. Moreover, the collaborative work involving Terminology and ontologies has led to the emergence of new theoretical perspectives, namely to the notion of Ontoterminology, which aims to reconcile Terminology's linguistic and conceptual dimension whilst preserving their core identities. This can be particularly relevant in subject fields such as Medicine, where concept-oriented and ontology-based approaches have become the cornerstone of the most recent (bio)medical terminological resources, and where non-verbal concept representations play a key role. Due to the lack of specialised lexicographic resources in the field of endometriosis, this paper aims to present the MODE project, i.e. the Multimodal Online Dictionary of Endometriosis, a multilingual resource comprising several types of data, namely video articles, a new type of scholarly communication in Medicine. It is believed that introducing a medical lexicographic resource supported by ontoterminological principles and encompassing scientific video articles may constitute a relevant window of opportunity in the research field of Lexicography.

Tuesday 24 May, 10:10 – 10:30
Chairperson: Michael Rundell

Determining the characteristic vocabulary for a specialized dictionary using Word2vec and a directed crawler

Gregory Grefenstette and Lawrence Muchemi

Specialized dictionaries are used to understand concepts in specific domains, especially where those concepts are not part of the general vocabulary, or having meanings that differ from ordinary languages. The first step in creating a specialized dictionary involves detecting the characteristic vocabulary of the domain in question. Classical methods for detecting this vocabulary involve gathering a domain corpus, calculating statistics on the terms found there, and then comparing these statistics to a background or general language corpus. Terms which are found significantly more often in the specialized corpus than in the background corpus are candidates for the characteristic vocabulary of the domain. Here we present two tools, a directed crawler, and a distributional semantics package, that can be used together, circumventing the need of a background corpus. Both tools are available on the web.

Tuesday 24 May, 11:00 – 11:20
Chairperson: Jelena Kallas

Karp: Språkbanken's open lexical infrastructure

Malin Ahlberg, Lars Borin, Markus Forsberg, Olof Olsson, Anne Schumacher and Jonatan Uppström

Karp is the open lexical infrastructure of Språkbanken (the Swedish Language Bank). As of today, there are 25+, mostly Swedish, lexical resources available in Karp, including modern lexicons designed for LT use, as well as older digitized dictionaries. Most resources, including the historical ones, have been at least partially linked to a pivot resource, SALDO, defining a connected network of Swedish lexical information. There are also multi-lingual resources representing more than 30 languages, as well as a lexicon for the ideographic writing system Bliss. Karp is being developed in collaboration with Swe-Clarin, and we pay close attention to its standards and best practices. Karp has been designed to support the creation and development of lexical resources. There are three main components: a REST-based web service, a graphical user interface, and an authentication server for managing user access. Users can add, update and remove entries, and a revision history is kept for each resource. A resource may have a group of authorized editors, but the system also allows for unauthorized users to give suggestions that can later be approved by editors. Karp provides user support during editing, such as feedback on the formatting, the compliance to a standard, or similar. The editing functionality in Karp has been a central component in several projects, among them are the Swedish Framenet++ (Ahlberg et al., 2014) and the Swedish Constructicon (Lyngfelt et al., 2014).

Tuesday 24 May, 11:20 – 11:40
Chairperson: Jelena Kallas

Semi-automatic compilation of a very large multiword expression dictionary for Bulgarian

Ivelina Stoyanova, Svetla Koeva, Maria Todorova and Svetlozara Leseva

The paper presents a very large dictionary of multiword expressions in Bulgarian which is compiled semi-automatically. We outline the main features of Bulgarian MWEs and their classification based on morphosyntactic, structural and semantic criteria. Further, we discuss the multi-layered organisation of the Dictionary and the components of the description of the MWEs, as well as the links to other lexicographic and general resources. Finally, we present the semi-automatic procedures for the compilation of the MWE entries. The work on the Dictionary is ongoing, aimed at extending its contents in terms of adding new entries and new layers of description, as well as at improving the quality of the resource.

Tuesday 24 May, 11:40 – 12:00
Chairperson: Jelena Kallas

CombiNet: Italian word combinations in an online lexicographic tool

Raffaele Simone and Valentina Piunno

This paper introduces CombiNet dictionary, an on line corpus-based lexicographic tool representing combinatorial properties of Italian lexemes, developed by Roma Tre University, University of Pisa and University of Bologna. The lexicographic layout of CombiNet is designed to include different sets of information, such as i) syntactic configurations and ii) syntactic function of word combinations, iii) degree of lexical variation associated with specific types of multiword units. In fact, CombiNet records word combinations showing different degrees of lexicalizations and paradigmatic variability, which is a novelty in lexicography. This investigation intends to tackle several issues associated with CombiNet, and in particular it aims at a) showing procedures and methods used to create and compile CombiNet's entries, b) describing particular types of combinatorial phenomena emerged from the analysis of corpus-based data, c) illustrating the lexicographic layout that has been elaborated for word combinations representation, d) describing the advanced research tool CombiNet is equipped with, a useful device for lexicographic investigations as well as for lexicological analysis.

Tuesday 24 May, 12:00 – 12:20
Chairperson: Jelena Kallas

GDEX for Japanese: Automatic extraction of good dictionary example candidates

Irena Srdanović and Iztok Kosem

The GDEX tool, devised to assist lexicographers in identifying good dictionary examples, was initially created for the English language (Kilgarriff et al., 2008) and proved very useful in various dictionary projects (c.f. Rundell & Kilgarriff, 2011). Later on, GDEX configurations were developed for Slovene (Kosem et al., 2011, 2013) and other languages. This paper employs similar methods to design GDEX for Japanese in order to extract good example candidates from Japanese language corpora available inside the Sketch Engine system. Criteria and parameters, which were adapted to Japanese language needs, were based on the configuration for Slovene as well as the default language independent configuration available in the Sketch Engine. A number of different configurations were devised and compared in order to identify optimal values for good example identification. The paper also explores a language-learner oriented approach to good example extraction by taking into account different difficulty levels of lexemes based on the Japanese Language Proficiency Test list of words and levels. For this purposes, additional configurations were devised, which are tailored to individual levels and thus useful for language learners and lexicographers of learner's dictionaries.

Tuesday 24 May, 12:20 – 12:40
Chairperson: Jelena Kallas

Towards a corpus-based valency lexicon of Czech nouns

Jana Klímová, Veronika Kolářová and Anna Vernerová

Corpus-based Valency Lexicon of Czech Nouns is a starting project picking up the threads of our previous work on nominal valency. It builds upon solid theoretical foundations of the theory of valency developed within the Functional Generative Description. In this paper, we describe the ways of treating valency of nouns in a modern corpus-based lexicon, available as machine readable data in a format suitable for NLP applications, and report on the limitations that the most commonly used corpus interfaces provide to the research of nominal valency. The linguistic material is extracted from the Prague Dependency Treebank, the synchronic written part of the Czech National Corpus, and Araneum Bohemicum. We will utilize lexicographic software and partially also data developed for the valency lexicon PDT-Vallex but the treatment of entries will be more exhaustive, for example, in the coverage of senses and in the semantic classification added to selected lexical units (meanings). The main criteria for including nouns in the lexicon will be semantic class membership and the complexity of valency patterns. Valency of nouns will be captured in the form of valency frames, enumeration of all possible combinations of adnominal participants, and corpus examples.

Tuesday 24 May, 14:20 – 14:40
Chairperson: Paul Cook

Verb argument pairing in a Czech-English parallel treebank

Jan Hajič, Eva Fučíková, Jana Šindlerová and Zdeňka Urešová

We describe CzEngVallex, a bilingual Czech-English valency lexicon which aligns verbal valency frames and their arguments. It is based on a parallel Czech-English corpus, the Prague Czech-English Dependency Treebank, where for each occurrence of a verb a reference to the underlying Czech and English valency lexicons is explicitly recorded. CzEngVallex lexicon pairs the entries (verb senses) of these two lexicons, and allows for detailed studies of verb valency and argument structure in translation. While some related studies have already been published on certain phenomena, we concentrate here on basic statistics, showing that the variability of verb argument mapping between verbs in the two languages is richer than it might seem and than the perception from the studies published so far might have been.

Tuesday 24 May, 14:40 – 15:00
Chairperson: Paul Cook

The role of computational Zulu verb morphology in multilingual lexicographic applications

Sonja Bosch and Laurette Pretorius

Performing cross-lingual natural language processing and developing multilingual lexicographic applications for languages with complex agglutinative morphology pose specific challenges that are aggravated when such languages are also under-resourced. In this paper, Zulu, an under-resourced language spoken in Southern Africa, is considered. The verb is the most complex word category in Zulu. Due to the agglutinative nature of Zulu morphology, limited information can be computationally extracted from running Zulu text without the support of sufficiently reliable computational morphological analysis by means of which the essential meanings of, amongst others, verbs can be exposed. The central research question that is addressed in this paper is as follows: How could ZulMorph (http://gama.unisa.ac.za/demo/demo/ZulMorph), a finite state morphological analyser for Zulu, be employed to support multilingual lexicography and cross-lingual natural language processing applications, with specific reference to Zulu verbs?

Tuesday 24 May, 15:00 – 15:20
Chairperson: Paul Cook

Toward a global online living dictionary: A model and obstacles for big linguistic data

Martin Benjamin

Having labelled itself the “Global Online Living Dictionary” (GOLD), this paper explores how Kamusi’s approach to multilingual lexicography is designed to fulfil that mission. This paper discusses recent under-the-hood work intended to produce an online compendium of linguistic data that satisfies many contemporary lexicographic concerns, working through hurdles to develop implementable systems for a viable lexicographic data resource for HLT knowledge and NLP uses across hundreds of languages.

Tuesday 24 May, 15:20 – 15:40
Chairperson: Paul Cook

Linking and disambiguating Swadesh lists

Luís Morgado da Costa, Francis Bond and František Kratochvíl

In this paper we describe two main contributions in the fields of lexicography and Linked Open Data: a human corrected disambiguation, using the Princeton Wordnet's sense inventory (PWN, Fellbaum, 1998), of Swadesh lists maintained in the Internet Archive by the Rosetta Project, and the distribution of this data through an expansion of the Open Multilingual Wordnet (OMW, Bond and Foster, 2013). The task of disambiguating word lists isn't always a straightforward task. The PWN is a vast resource with many fine-grained senses, and word lists often fail to help resolve the inherent ambiguity of words. In this work we describe the corner cases of this disambiguation and, when necessary, motivate our choice over other possible senses. We take the results of this work as a great example of the benefits of sharing linguistic data under open licenses, and will continue linking other openly available data. All the data will be released in future OMW releases, and we will encourage the community to contribute in correcting and adding to the data made available.

Tuesday 24 May, 15:40 – 16:00
Chairperson: Paul Cook

DefExt: A semi-supervised definition extraction tool

Luis Espinosa Anke, Roberto Carlini, Horacio Saggion and Francesco Ronzano

We present DEFEXT, an easy to use semi supervised Definition Extraction Tool. DEFEXT is designed to extract from a target corpus those textual fragments where a term is explicitly mentioned together with its core features, i.e. its definition. It works on the back of a Conditional Random Fields based sequential labeling algorithm and a bootstrapping approach. Bootstrapping enables the model to gradually become more aware of the idiosyncrasies of the target corpus. In this paper we describe the main components of the toolkit as well as experimental results stemming from both automatic and manual evaluation. We release DEFEXT as open source along with the necessary files to run it in any Unix machine. We also provide access to training and test data for immediate use.

Tuesday 24 May, 16:30 – 16:50
Chairperson: Laurette Pretorius

Modelling multilingual lexicographic resources for the web of data: The K Dictionaries case

Julia Bosque-Gil, Jorge Gracia, Elena Montiel-Ponsoda and Guadalupe Aguado-de-Cea

Lexicographic resources can highly benefit from Semantic Web technologies, specifically, linked data technologies, since such resources cannot only become easy to access and query, but also easy to share and link to resources that contain complementary information, contributing to the creation of a huge graph of interlinked lexical and linguistic resources. In this paper, we present the methodology we have followed for the transformation of a lexicographic resource, namely, the Spanish dataset of K Dictionaries's Global series, from its proprietary XML format to RDF, according to the lemon-ontolex model, a de-facto standard for representing lexical information in the Web of Data. We describe in detail the original resource, the design decisions taken for the transformation process, the model chosen for the representation of the dataset, as well as the extensions made to the model to accommodate specific modelling needs of the original source. The core of the representation model is described in detail in order to illustrate the issues encountered and how they have been solved in this first prototype, which could serve to lay the foundations for future transformations.

Tuesday 24 May, 16:50 – 17:10
Chairperson: Laurette Pretorius

Creating seed lexicons for under-resourced languages

Ivett Benyeda, Péter Koczka and Tamás Váradi

In this paper we present methods of creating seed dictionaries for an under-resourced language, Udmurt, paired with four thriving languages. As reliable machine readable dictionaries do not exist in desired quantities this step is crucial to enable further NLP tasks, as seed dictionaries can be considered the first connecting element between two sets of texts. For the language pairs discussed in this paper, detailed description will be given of various methods of translation pair extraction, namely Wik2Dict, triangulation, Wikipedia article title pair extraction and handling the problematic aspects, such as multiword expressions (MWUs) among others. After merging the created dictionaries we were able to create seed dictionaries for all language pairs with approximately a thousand entries, which will be used for sentence alignment in future steps and thus will aid the extraction of larger dictionaries.

Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem

24 May 2016

ABSTRACTS

Editors:

Georg Rehm, Aljoscha Burchardt, Ondřej Bojar, Christian Dugast, Marcello Federico, Josef van Genabith, Barry Haddow, Jan Hajič, Kim Harris, Philipp Koehn, Matteo Negri, Martin Popel, Lucia Specia, Marco Turchi, Hans Uszkoreit

Workshop Programme

09.00 – 09.10 – Welcome – introduction – context

Session 1: Tools, Methods and Resources for Research

09.10 – 09.30 – Julia Ive, Aurélien Max, François Yvon, Philippe Ravaud, *Diagnosing High-Quality Statistical Machine Translation Using Traces of Post-Edition Operations*

09.30 – 09.50 – Ondřej Bojar, Filip Děchtěrenko, Maria Zelenina, *A Pilot Eye-Tracking Study of WMT-Style Ranking Evaluation*

09.50 – 10.00 – Anabela Barreiro, Francisco Raposo, Tiago Luís, CLUE-Aligner: *An Alignment Tool to Annotate Pairs of Paraphrastic and Translation Units (short presentation)*

10.00 – 10.10 – Zijian Győző Yang, László János Laki, Borbála Siklósi, *HuQ: An English-Hungarian Corpus for Quality Estimation (short presentation)*

10.10 – 10.30 – Discussion of the papers presented in Session 1

10.30 – 11.00 – Coffee break

Session 2: Shared Tasks

11.00 – 11.20 – Ondřej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, Lucia Specia, *Ten Years of WMT Evaluation Campaigns: Lessons Learnt*

11.20 – 11.40 – Luisa Bentivogli, Marcello Federico, Sebastian Stüker, Mauro Cettolo, Jan Niehues, *The IWSLT Evaluation Campaign: Challenges, Achievements, Future Directions*

11.40 – 12.00 – Discussion of the papers presented in Session 2

Session 3: Evaluation Tools and Metrics (part A)

12.00 – 12.20 – Katrin Marheinecke, *Can Quality Metrics Become the Drivers for Machine Translation Uptake? An Industry Perspective*

12.20 – 12.40 – Kim Harris, Aljoscha Burchardt, Georg Rehm, Lucia Specia, *Technology Landscape for Quality Evaluation: Combining the Needs of Research and Industry*

12.40 – 13.00 – Eleftherios Avramidis, *Interoperability in MT Quality Estimation or wrapping useful stuff in various ways*

13.00 – 14.00 – Lunch break

Session 3: Evaluation Tools and Metrics (part B)

14.00 – 14.20 – Arle Lommel, *Blues for BLEU: Reconsidering the Validity of Reference-Based MT Evaluation*

14.20 – 14.40 – Roman Sudarikov, Martin Popel, Ondřej Bojar, Aljoscha Burchardt, Ondřej Klejch, *Using MT-ComparEval*

14.40 – 15.00 – Michal Tyszkowski, Dorota Szaszko, *CMT: Predictive Machine Translation Quality Evaluation Metric*

15.00 – 15.20 – Aljoscha Burchardt, Kim Harris, Georg Rehm, Hans Uszkoreit, *Towards a Systematic and Human-Informed Paradigm for High-Quality Machine Translation*

15.20 – 16.00 – Discussion of the papers presented in Session 3 (part A, partB)

16.00 – 16.30 – Coffee break

16.30 – 17.30 – Summary – final discussion – next steps: towards an integrated ecosystem?

17.30 – End of workshop

Organising Committee

Ondřej Bojar	Charles University in Prague, Czech Republic
Aljoscha Burchardt	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany*
Christian Dugast	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
Marcello Federico	Fondazione Bruno Kessler (FBK), Italy
Josef van Genabith	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
Barry Haddow	University of Edinburgh, UK
Jan Hajič	Charles University in Prague, Czech Republic
Kim Harris	text&form, Germany
Philipp Koehn	Johns Hopkins University, USA, and University of Edinburgh, UK
Matteo Negri	Fondazione Bruno Kessler (FBK), Italy
Martin Popel	Charles University in Prague, Czech Republic
Georg Rehm	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany*
Lucia Specia	University of Sheffield, UK
Marco Turchi	Fondazione Bruno Kessler (FBK), Italy
Hans Uszkoreit	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany

* main editors and chairs of the Organising Committee

Programme Committee

Nora Aranberri	University of the Basque Country, Spain
Ondřej Bojar	Charles University in Prague, Czech Republic
Aljoscha Burchardt	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
Christian Dugast	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
Marcello Federico	Fondazione Bruno Kessler (FBK), Italy
Christian Federmann	Microsoft, USA
Rosa Gaudio	Higher Functions, Portugal
Josef van Genabith	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
Barry Haddow	University of Edinburgh, UK
Jan Hajič	Charles University in Prague, Czech Republic
Kim Harris	text&form, Germany
Matthias Heyn	SDL, Belgium
Philipp Koehn	Johns Hopkins University, USA, and University of Edinburgh, UK
Christian Lieske	SAP, Germany
Lena Marg	Welocalize, UK
Katrin Marheinecke	text&form, Germany
Matteo Negri	Fondazione Bruno Kessler (FBK), Italy
Martin Popel	Charles University in Prague, Czech Republic
Jörg Porsiel	Volkswagen AG, Germany
Georg Rehm	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany
Rubén Rodriguez de la Fuente	PayPal, Spain
Lucia Specia	University of Sheffield, UK
Marco Turchi	Fondazione Bruno Kessler (FBK), Italy
Hans Uszkoreit	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany

Preface

Current approaches to Machine Translation (MT) or professional translation evaluation, both automatic and manual, are characterised by a high degree of fragmentation, heterogeneity and a lack of interoperability between methods, tools and data sets. As a consequence, it is difficult to reproduce, interpret, and compare evaluation results. In an attempt to address this issue, the main objective of this workshop is to bring together researchers working on translation evaluation and practitioners (translators, users of MT, language service providers etc.).

This workshop takes an in-depth look at an area of ever-increasing importance. Two clear trends have emerged over the past several years. The first trend involves standardising evaluations in research through large shared tasks in which actual translations are compared to reference translations using automatic metrics or human ranking. The second trend focuses on achieving high quality (HQ) translations with the help of increasingly complex data sets that contain many levels of annotation based on sophisticated quality metrics – often organised in the context of smaller shared tasks. In industry, we also observe an increased interest in workflows for HQ outbound translation that combine Translation Memories, MT, and post-editing. In stark contrast to this trend to quality translation and ist inherent overall approach and complexity, the data and tooling landscapes remain rather heterogeneous, uncoordinated and not interoperable.

The event brings together researchers, users and providers of tools, and users and providers of manual and automatic translation evaluation methodologies. We want to initiate a dialogue and discuss whether the current approach involving a diverse and heterogeneous and distributed set of data, tools, scripts, and evaluation methodologies is appropriate enough or if the community should, instead, collaborate towards building an integrated ecosystem that provides better and more sustainable access to data sets, evaluation workflows, tools, approaches, and metrics that support processes such as annotations, quality comparisons and post-editing.

The workshop is meant to stimulate a dialogue about the commonalities, similarities and differences of the existing solutions in the three areas (1) tools, (2) methodologies, (3) data sets. A key question concerns the high level of flexibility and lack of interoperability of heterogeneous approaches, while a homogeneous approach would provide less flexibility but higher interoperability and thus allow, e.g., integrated research by means of an MT app store (cf. The *Translingual Cloud* anticipated in the META-NET Strategic Research Agenda). How much flexibility and interoperability does the translation community need? How much does it want? How can communication and collaboration between industry and research be intensified?

We hope that the papers presented and discussed at the workshop provide at least partial answers on these, and other, crucial questions around the complex and interdisciplinary topic of evaluating translations, either produced by machines or by human experts.

G. Rehm, A. Burchardt, O. Bojar, C. Dugast, M. Federico, J. van Genabith, B. Haddow, J. Hajič, K. Harris, P. Koehn, M. Negri, M. Popel, L. Specia, M. Turchi, H. Uszkoreit May 2016

Abstracts

Session 1: Tools, Methods and Resources for Research

Tuesday, 24 May 2016, 09:00 – 10:30

Diagnosing High-Quality Statistical Machine Translation Using Traces of Post-Edition Operations

Julia Ive, Aurélien Max, François Yvon, Philippe Ravaud

This paper proposes a fine-grained flexible analysis methodology to reveal the residual difficulties of a high-quality Statistical Machine Translation (SMT) system. This proposal is motivated by the fact that the traditional automated metrics are not enough informative to indicate the nature and reasons of those residual difficulties. Their resolution is however a key point towards improving the high-quality output. The novelty of our approach consists in diagnosing Machine Translation (MT) performance by making a connection between errors, the characteristics of source sentences and some internal parameters of the system, using traces of Post-Edition (PE) operations as well as Quality Estimation (QE) techniques. Our methodology is illustrated on a SMT system adapted to the medical domain, based on a high quality English-French parallel corpus of Cochrane systematic review abstracts. Our experimental results show that the main difficulties that the system faces are in the domains of term precision and source language syntactic and stylistic peculiarities. We furthermore provide general information regarding the corpus structure and its specificities, including internal stylistic varieties characteristic of this sub-genre.

A Pilot Eye-Tracking Study of WMT-Style Ranking Evaluation

Ondřej Bojar, Filip Děchtěrenko, Maria Zelenina

The shared translation task of the Workshop of Statistical Machine Translation (WMT) is one of the key annual events of the field. Participating machine translation systems in WMT translation task are manually evaluated by relatively ranking five candidate translations of a given sentence. This style of evaluation has been used since 2007 with some discussion on interpreting the collected judgements but virtually no insight into what the annotators are actually doing. The scoring task is relatively cognitively demanding and many scoring strategies are possible, influencing the reliability of the final judgements. In this paper, we describe our first steps towards explaining the scoring task: we run the scoring under an eye-tracker and monitor what the annotators do. At the current stage, our results are more of a proof-of-concept, testing the feasibility of eye tracking for the analysis of such a complex MT evaluation setup.

CLUE-Aligner: An Alignment Tool to Annotate Pairs of Paraphrastic and Translation Units

Anabela Barreiro, Francisco Raposo, Tiago Luís

Currently available alignment tools and procedures for marking-up alignments overlook non-contiguous multiword units for being too complex within the bounds of the proposed alignment methodologies. This paper presents the CLUE-Aligner (Cross-Language Unit Elicitation Aligner), a web alignment tool designed for manual annotation of pairs of paraphrastic and translation units, representing both contiguous and non-contiguous multiwords and phrasal expressions found in monolingual or bilingual parallel sentences. Non-contiguous block alignments are necessary to express alignments between multiwords or phrases, which contain insertions, i.e., words that are not part of the multiword unit or phrase. CLUE-Aligner also allows the alignment of smaller individual or multiword units inside non-contiguous multiword units. The interactive web application was developed under the scope of the eSPERTo project, which aims to build a linguistically enhanced paraphrasing system. However, a tool for manual annotation of alignment and for visualization of automatic phrase alignment can prove useful in human and machine translation evaluation.

HuQ: An English-Hungarian Corpus for Quality Estimation

Zijian Győző Yang, László János Laki, Borbála Siklósi

Quality estimation for machine translation is an important task. The standard automatic evaluation methods that use reference translations cannot perform the evaluation task well enough. These methods produce low correlation with human evaluation for English-Hungarian. Quality estimation is a new approach to solve this problem. This method is a prediction task estimating the quality of translations for which features are extracted from only the source and translated sentences. Quality estimation systems have not been implemented for Hungarian before, thus there is no such training corpus either. In this study, we created a dataset to build quality estimation models for English-Hungarian. We also did experiments to optimize the quality estimation system for Hungarian. In the optimization task we did research in the field of feature engineering and feature selection. We created optimized feature sets, which produced better results than the baseline feature set.

Session 2: Shared Tasks

Tuesday, 24 May 2016, 11:00 – 12:00

Ten Years of WMT Evaluation Campaigns: Lessons Learnt

Ondřej Bojar, Christian Federmann, Barry Haddow, Philipp Koehn, Matt Post, Lucia Specia

The WMT evaluation campaign (http://www.statmt.org/wmt16) has been run annually since 2006. It is a collection of shared tasks related to machine translation, in which researchers compare their techniques against those of others in the field. The longest running task in the campaign is the translation task, where participants translate a common test set with their MT systems. In addition to the translation task, we have also included shared tasks on evaluation: both on automatic metrics (since 2008), which compare the reference to the MT system output, and on quality estimation (since 2012), where system output is evaluated without a reference. An important component of WMT has always been the manual evaluation, wherein human annotators are used to produce the official ranking of the systems in each translation task. This reflects the belief of the WMT organizers that human judgement should be the ultimate arbiter of MT quality. Over the years, we have experimented with different methods of improving the reliability, efficiency and discriminatory power of these judgements. In this paper we report on our experiences in running this evaluation campaign, the current state of the art in MT evaluation (both human and automatic), and our plans for future editions of WMT.

The IWSLT Evaluation Campaign: Challenges, Achievements, Future Directions

Luisa Bentivogli, Marcello Federico, Sebastian Stüker, Mauro Cettolo, Jan Niehues

Evaluation campaigns are the most successful modality for promoting the assessment of the state of the art of a field on a specific task. Within the field of Machine Translation (MT), the International Workshop on Spoken Language Translation (IWSLT) is a yearly scientific workshop, associated with an open evaluation campaign on spoken language translation. The IWSLT campaign, which is the only one addressing speech translation, started in 2004 and will feature its 13th installment in 2016. Since its beginning, the campaign attracted around 70 different participating teams from all over the world. In this paper we present the main characteristics of the tasks offered within IWSLT, as well as the evaluation framework adopted and the data made available to the research community. We also analyse and discuss the progress made by the systems along the years for the most addressed and long-standing tasks and we share ideas about new challenging data and interesting application scenarios to test the utility of MT systems in real tasks.

Session 3: Evaluation Tools and Metrics (part A)

Tuesday, 24 May 2016, 12:00 – 13:00

Can Quality Metrics Become the Drivers of Machine Translation Uptake? An Industry Perspective

Katrin Marheinecke

Language service providers (LSPs) who want to make use of Machine Translation (MT) have to fight on several fronts. The skepticism within the language industry is still very high: End-customers worry about paying too much for translations that no human has interfered with. Translators refuse to get involved in post-editing activities because they fear that MT will take away their actual work, rendering themselves useless on the long run. And competitors try to outperform each other by either spoiling the market with low-quality MT offered on dumping rates or declining MT altogether for being inappropriate for commercial usage. This paper seeks to show that well-defined quality metrics can help all stakeholders of the translation market to specify adequate benchmarks for the desired translation quality, to use an agreed-upon consistent mark-up and to evaluate translation quality – MT and human translation output alike – accordingly. As a by-product, this professionally error-annotated MT output will help researchers to further improve MT quality, which in turn will help to make this technology more popular in the industry.

Technology Landscape for Quality Evaluation: Combining the Needs of Research and Industry

Kim Harris, Aljoscha Burchardt, Georg Rehm, Lucia Specia

Translation quality evaluation (QE) has gained significant uptake in recent years, in particular in light of increased demand for automated translation workflows and machine translation. Despite the need for innovative and forward-looking quality evaluation solutions, the technology landscape remains highly fragmented and the two major constituencies in need of collaborative and ground-breaking technology are still very divided. This paper will demonstrate that closer cooperation between users of QE technology in research and industry to create a holistic but highly adaptable environment for all aspects of the translation improvement process, most significantly quality evaluation, can lead the way to novel and ground-breaking achievements in accelerated improvement in machine translation results.

Interoperability in MT Quality Estimation or wrapping useful stuff in various ways

Eleftherios Avramidis

The situation on the interoperability of Natural Language Processing software is outlined through a use-case on Quality Estimation of Machine Translation output. The focus is on the development efforts for the QUALITATIVE tool, so that it integrates a multitude of state-of-the-art external tools into one single Python program, through an interoperable framework. The presentation includes 9 approaches taken to connect 25 external components, developed in various programming languages. The conclusion is that the current landscape lacks important interoperability principles and that developers should be encouraged to equip their programs with some of the standard interaction interfaces.

Session 3: Evaluation Tools and Metrics (part B)

Tuesday, 24 May 2016, 14:00 – 16:00

Blues for BLEU: Reconsidering the Validity of Reference-Based MT Evaluation

Arle Lommel

This article describes experiments a set of experiments designed to test whether reference-based machine translation evaluation methods (represented by BLEU) (a) measure translation “quality” and (b) whether the scores they generate are reliable as a measure for systems (rather than for particular texts). It considers these questions via three methods. First, it examines the impact of changing reference translations and using them in combination on BLEU scores. Second, it examines the internal consistency of BLEU scores, the extent to which reference-based scores for a part of a text represent the score of the whole. Third, it applies BLEU to human translation to determine whether BLEU can reliably distinguish human translation from MT output. The results of these experiments, conducted on a Chinese>English news corpus with eleven human reference translations, bring the validity of BLEU as a measure of translation quality into question and suggest that the score differences cited in a considerable body of MT literature are likely to be unreliable indicators of system performance due to an inherent imprecision in reference-based methods. Although previous research has found that human quality judgments largely correlate with BLEU, this study suggests that the correlation is an artefact of experimental design rather than an indicator of validity.

Using MT-ComparEval

Roman Sudarikov, Martin Popel, Ondřej Bojar, Aljoscha Burchardt, Ondřej Klejch

The paper showcases the MT-ComparEval tool for qualitative evaluation of machine translation (MT). MT-ComparEval is an open-source tool that has been designed in order to help MT developers by providing a graphical user interface that allows the comparison and evaluation of different MT engines/experiments and settings. The tool implements several measures that represent the current best practice of automatic evaluation. It also provides guidance in the targeted inspection of examples that show a certain behavior in terms of n-gram similarity/dissimilarity with alternative translations or the reference translation. In this paper, we provide an applied, “hands-on” perspective on the actual usage of MT-ComparEval. In a case study, we use it to compare and analyze several systems submitted to the WMT 2015 shared task.

CMT: Predictive Machine Translation Quality Evaluation Metric

Michal Tyszkowski, Dorota Szaszko

Machine Translation quality is evaluated using metrics that utilize human translations as reference materials. This means that the existing methods do not allow to predict the quality of Machine Translation if no human translations of the same material exist. To use Machine Translation in the translation industry, it is essential to have a metric that allows to evaluate the quality of a newly created Machine Translation in order to decide whether or not it can increase productivity. As a translation company that uses Statistical Machine Translation to generate translation memories for many projects, we decided to develop a metric that can predict its usability on a project basis. This metric, called CMT, is a combination of human assessment and statistics comparable to existing metrics. Our investigations showed that the mean difference between CMT and BLEU is 9.10, and between CMT and METEOR it is 9.69, so it correlates with the existing metrics in more than 90%. CMT is very easy to use and allows to evaluate each translation memory, regardless of its size, in 5 minutes without any reference material.

Towards a Systematic and Human-Informed Paradigm for High-Quality Machine Translation

Aljoscha Burchardt, Kim Harris, Georg Rehm, Hans Uszkoreit

Since the advent of modern statistical machine translation (SMT), much progress in system performance has been achieved that went hand-in-hand with ever more sophisticated mathematical models and methods. Numerous small improvements have been reported whose lasting effects are hard to judge, especially when they are combined with other newly proposed modifications of the basic models. Often the measured enhancements are hardly visible with the naked eye and two performance advances of the same measured magnitude are difficult to compare in their qualitative effects. We sense a strong need for a paradigm in MT research and development (R&D), that pays more attention to the subject matter, i.e., translation, and that analytically concentrates on the many different challenges for quality translation. The approach we propose utilizes the knowledge and experience of professional translators throughout the entire R&D cycle. It focuses on empirically confirmed quality barriers with the help of standardised error metrics that are supported by a system of interoperable methods and tools and are shared by research and translation business.

The 2nd Workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media

Date 24 May 2016

ABSTRACTS

Editors:

Hend Al-Khalifa, Abdulmohsen Al-Thubaity, Walid Magdy and

Kareem Darwish

Workshop Programme

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

09:20 – 10:30 – Session 1 (Keynote speech)

Nizar Habash, *Computational Processing of Arabic Dialects: Challenges, Advances and Future Directions*

10:30 – 11:00 Coffee break

10:30 – 13:00 – Session 2

Soumia Bougrine, Hadda Cherroun, Djelloul Ziadi, Abdallah Lakhdari and Aicha Chorana, *Toward a rich Arabic Speech Parallel Corpus for Algerian sub-Dialects*

Maha Alamri and William John Teahan, *Towards a New Arabic Corpus of Dyslexic Texts*

Ossama Obeid, Houda Bouamor, Wajdi Zaghouani, Mahmoud Ghoneim, Abdelati Hawwari, Mona Diab and Kemal Oflazer, *MANDIAC: A Web-based Annotation System For Manual Arabic Diacritization*

Wajdi Zaghouani and Dana Awad, *Toward an Arabic Punctuated Corpus: Annotation Guidelines and Evaluation*

Muhammad Abdul-Mageed, Hassan Alhuzali, Dua'a Abu-Elhij'a and Mona Diab, *DINA: A Multi-Dialect Dataset for Arabic Emotion Analysis*

Nora Al-Twairesh, Mawaheb Al-Tuwaijri, Afnan Al-Moammar and Sarah Al-Humoud, *Arabic Spam Detection in Twitter*

Workshop Organizers

Hend Al-Khalifa	King Saud University, KSA
Abdulmohsen Al-Thubaity	King Abdul Aziz City for Science and Technology, KSA
Walid Magdy	Qatar Computing Research Institute, Qatar
Kareem Darwish	Qatar Computing Research Institute, Qatar

Workshop Programme Committee

Abdullah Alfaifi	Imam University, KSA
Abdulrhman Almuhareb	King Abdul Aziz City for Science and Technology, KSA
Abeer ALDayel	King Saud University, KSA
Ahmed Abdelali	Qatar Computing Research Institute, Qatar
Areeb AlOwisheq	Imam University, KSA
Auhood Alfaries	King Saud University, KSA
Hamdy Mubarak	Qatar Computing Research Institute, Qatar
Hazem Hajj	American University of Beirut, Lebanon
Hind Al-Otaibi	King Saud University, KSA
Houda Bouamor	Carnegie Mellon University, Qatar
Khurshid Ahmad	Trinity College Dublin, Ireland
Maha Alrabiah	Imam University, KSA
Mohammad Alkanhal	King Abdul Aziz City for Science and Technology, KSA
Mohsen Rashwan	Cairo University, Egypt
Mona Diab	George Washington University, USA
Muhammad M. Abdul-Mageed	Indiana University, USA
Nizar Habash	New York University Abu Dhabi, UAE
Nora Al-Twairesh	King Saud University, KSA
Nouf Al-Shenaifi	King Saud University, KSA
Tamer Elsayed	Qatar University, Qatar
Wajdi Zaghouani	Carnegie Mellon University in Qatar, Qatar

Preface

Given the success of our first Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools in LREC 2014 where three of the presented papers received 15 citations up to now. The second workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools (OSACT2) with special emphasis on Arabic social media text processing and applications aims to encourage researchers and developers to foster the utilization of freely available Arabic corpora and open source Arabic corpora processing tools and help in highlighting the drawbacks of these resources and discuss techniques and approaches on how to improve them.

OSACT2 had an acceptance rate of 55%, where we received 11 papers from which 6 papers were accepted. We believe the accepted papers are high quality and present mixture of interesting topics. Three papers are about corpus development and annotation guidelines for different domains such as speech and dyslexic texts, two papers about spam detection and emotion analysis, and finally, a paper presenting a web-based system for manual annotation of Arabic diacritization.

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 Dr. Nizar Habash for accepting to give the workshop keynote talk, to the members of the program committee who did an excellent job in reviewing the submitted papers, and to the LREC organizers. Last but not least we would like to thank our authors and the participants of the workshop.

Hend Al-Khalifa, Abdulmohsen Al-Thubaity, Walid Wagdy and Kareem Darwish
Portorož (Slovenia), 2016

Session 1

Tuesday 24 May, 9:20 – 10:30

Computational Processing of Arabic Dialects: Challenges, Advances and Future Directions

Nizar Habash

Abstract

The Arabic language consists of a number of variants among which Modern Standard Arabic (MSA) has a special status as the formal, mostly written, standard of the media, culture and education across the Arab World. The other variants are informal, mostly spoken, dialects that are the languages of communication of daily life. Most of the natural language processing resources and research in Arabic have focused on MSA. However, recently, more and more research is targeting Arabic dialects. In this talk, we present the main challenges of processing Arabic dialects, and discuss common solution paradigms, current advances, and future directions.

Session 2

Tuesday 24 May, 11:00 – 13:00

Toward a rich Arabic Speech Parallel Corpus for Algerian sub-Dialects

Soumia Bougrine, Hadda Cherroun, Djelloul Ziadi, Abdallah Lakhdari and Aicha Chorana

Abstract

Speech datasets and corpora are crucial for both developing and evaluating accurate Natural Language Processing systems. While Modern Standard Arabic has received more attention, dialects are drastically underestimated, even they are the most used in our daily life and the social media, recently. In this paper, we present the methodology of building an Arabic Speech Corpus for Algerian dialects, and the preliminary version of that dataset of dialectal Arabic speeches uttered by Algerian native speakers selected from different Algeria's departments. In fact, by means of a direct recording way, we have taken into account numerous aspects that foster the richness of the corpus and that provide a representation of phonetic, prosodic and orthographic varieties of Algerian dialects. Among these considerations, we have designed a rich speech topics and content. The annotations provided are some useful information related to the speakers, time-aligned orthographic word transcription. Many potential uses can be considered such as speaker/dialect identification and computational linguistic for Algerian sub-dialects. In its preliminary version, our corpus encompasses 17 sub-dialects with 109 speakers and more than 6 K utterances.

Towards a New Arabic Corpus of Dyslexic Texts

Maha Alamri and William John Teahan

Abstract

This paper presents a detailed account of the preliminary work for the creation of a new Arabic corpus of dyslexic text. The analysis of errors found in the corpus revealed that there are four types of spelling errors made as a result of dyslexia in addition to four common spelling errors. The subsequent aim was to develop a spellchecker capable of automatically correcting the spelling mistakes of dyslexic writers in Arabic texts using statistical techniques. The purpose was to provide a tool to assist Arabic dyslexic writers. Some initial success was achieved in the automatic correction of dyslexic errors in Arabic text.

MANDIAC: A Web-based Annotation System For Manual Arabic Diacritization

Ossama Obeid, Houda Bouamor, Wajdi Zaghouani, Mahmoud Ghoneim, Abdelati Hawwari, Mona Diab and Kemal Oflazer

Abstract
In this paper, we introduce MANDIAC, a web-based annotation system designed for rapid manual diacritization of Standard Arabic text. To expedite the annotation process, the system provides annotators with a choice of automatically generated diacritization possibilities for each word. Our framework provides intuitive interfaces for annotating text and managing the diacritization annotation process. In this paper we describe the annotation and the administration interfaces as well as the back-end engine. Finally, we demonstrate that our system doubles the annotation speed compared to using a regular text editor.

Toward an Arabic Punctuated Corpus: Annotation Guidelines and Evaluation

Wajdi Zaghouani and Dana Awad

Abstract
We present our effort to build a large scale punctuated corpus for Arabic. We illustrate in details our punctuation annotation guidelines designed to improve the annotation work flow and the inter-annotator agreement. We summarize the guidelines created, discuss the annotation framework and show the Arabic punctuation peculiarities. Our guidelines were used by trained annotators and regular inter-annotator agreement measures were performed to ensure the annotation quality. We highlight the main difficulties related to the Arabic punctuation annotation that arose during this project.

DINA: A Multi-Dialect Dataset for Arabic Emotion Analysis

Muhammad Abdul-Mageed, Hassan Alhuzali, Dua'a Abu-Elhij'a and Mona Diab

Abstract
Although there has been a surge of research on sentiment analysis, less work has been done on the related task of emotion detection. Especially for the Arabic language, there is no literature that we know of for the computational treatment of emotion. This situation is due partially to lack of labelled data, a bottleneck that we seek to ease. In this work, we report efforts to acquire and annotate a multi-dialect dataset for Arabic emotion analysis.

Arabic Spam Detection in Twitter

Nora Al-Twairesh, Mawaheb Al-Tuwaijri, Afnan Al-Moammar and Sarah Al-Humoud

Abstract
Spam in Twitter has emerged due to the proliferation of this social network among users worldwide coupled with the ease of creating content. Having different characteristics than Web or mail spam, Twitter spam detection approaches have become a new research problem. This study aims to analyse the content of Saudi tweets to detect spam by developing both a rule-based approach that exploits a spam lexicon extracted from the tweets and a supervised learning approach that utilizes statistical methods based on the bag of words model and several features. The focus is on spam in trending hashtags in the Saudi Twittersphere since most of the spam in Saudi tweets is found in hashtags. The features used were identified through empirical analysis then applied in the

classification approaches developed. Both approaches showed comparable results in terms of performance measures reported reaching an average F-measure of 85% for the rule based approach and 91.6% for the supervised learning approach.

WILDRE3 – 3RDWorkshopon Indian Language Data: Resources and Evaluation

24th May 2016

ABSTRACTS

Editors:

Girish Nath Jha, Kalika Bali, Sobha L, Atul Kr. Ojha

Workshop Programme

24th May 2016

14:00 – 15:00hrs: Inaugural session

14:00 – 14:10 hrs – Welcome by Workshop Chairs
14:10 – 14:30 hrs – Inaugural Address
14:30 – 15:00hrs – Keynote Lecture

15:00 – 16:00 hrs – Paper Session I (Oral and Short Oral Presentation)
Chairperson: **Kalika Bali**

Oral Presentation

- Alok Parlikar, Sunayana Sitaram, Andrew Wilkinson and Alan W Black, *The Festvox Indic Frontend for Grapheme to Phoneme Conversion*
- Bornali Phukon, Biswajit Dev Sarma, Shakuntala Mahanta and S R M Prasanna, *Automatic Phonetic Alignment Tool Based on Hidden Markov Model as a Plug-in Tool of Praat for the Languages of Northeast India*

Short Oral Presentation

- Tehreem Waseem, Noorulain Ashraf, Shireen Gull and Sadaf Abdul Rauf, *Text Normalization in Urdu Text-to-Speech Synthesis System*
- Renu Singh, Atul Kr. Ojha and Girish Nath Jha, *Classification and Identification of Reduplicated Multi-Word Expressions in Hindi*
- Pattabhi RK Rao and Sobha Lalitha Devi, *Semantic Representation of Tamil Texts using Conceptual Graphs*
- Srishti Singh, *Valance Annotation of Hindi on Typecraft*
- Pitambar Behera, *Evaluation of SVM-based Automatic Parts of Speech Tagger for Odia*

16:00 – 16:30 hrs – Coffee break + Poster/Demo
Chairperson: **Massimo Moneglia/Lars Hellan**

- Divyanshu Bhardwaj and Amitava Das, *Part-of-Speech Tagging System for Maithili Tweets*
- Sharmin Muzaffar, Pitambar Behera and Girish Nath Jha, *Classification and Resolution of Linguistic Divergences in English-Urdu Machine Translation*
- Milind Shivolkar, Jyoti D. Pawar and S. Baskar, *Extractive based Email Summarization: An Unsupervised Hybrid approach using Graph Based Sentence Ranking and K-Means Clustering algorithm*
- Ashweta Fondekar, Jyoti D. Pawar and Ramdas Karmali, *Konkani SentiWordNet - Resource For Sentiment Analysis Using Supervised Learning Approach*
- Kumar Nripendra Pathak and Girish Nath Jha, *Verb Mapping: A Dilemma in Sanskrit-Hindi Machine Translation*
- Subhash Chandra, Bhupendra Kumar, Vivek Kumar and Sakshi, *Lexical Resources for Recognition, Analysis and Word Formation process for Sanskrit Morphology*
- Archana Tiwari, *Building a Statistical tagger for Sanskrit*
- Ritesh Kumar, Atul Kr. Ojha and Bornini Lahiri, *Developing annotated multimodal corpus for automatic recognition of verbal aggression in Hindi*
- Jalpa Zaladi and Caroline Gasperin, *Demo: SwiftKey Keyboard for Indian Languages*

- Shruti Rijhwani, Royal Sequeira, Monojit Choudhury and Kalika Bali, *Translating Code-Mixed Tweets: A Language Detection Based System*
- Sobha Lalitha Devi, Vijay Sundar Ram, Sindhuja Gopalan, Pattabhi RK Rao and Lakshmi S, *Demo Proposal: Tamil – Hindi Automatic Machine Translation – A detailed Description*
- Sobha Lalitha Devi, Pattabhi RK Rao, Vijay Sundar Ram and C.S Malarkodi, *A Demo Proposal: Tamil – English Cross Lingual Information Access (CLIA) System*

16.30 – 17:30 hrs – Paper Session II (Oral Presentation)
Chairperson: **Zygmunt Vetulani**

- Lars Borin, Shafqat Mumtaz Virk and Anju Saxena, *Towards a Big Data View on South Asian Linguistic Diversity*
- Atul Kr. Ojha, Srishti Singh, Pitambar Behera and Girish Nath Jha, *A Hybrid Chunker for Hindi and Indian English*
- Akshay Gadre, Rafiya Begum, Kalika Bali and Monojit Choudhury, *Machine Translating Code Mixed Text: Pain Points and Sweet Spots*
- Kavita Asnani and Jyoti D. Pawar, *Discovering Thematic Knowledge from Code-Mixed Chat Messages Using Topic Model*
- Pitambar Behera, Renu Singh and Girish Nath Jha, *Evaluation of Anuvadaksh (EILMT) English-Odia Machine-assisted Translation Tool*

17:30 – 18:10 hrs – Panel discussion
Coordinator: TBD
Panellists – TBD

18:10 – 18:25 hrs – Valedictory Address

18:25 – 18:30 hrs – Vote of Thanks

Workshop Organizers

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

Workshop Programme Committee

A G Ramakrishnan	I.I.Sc Bangalore
A Kumaran	Sri Foundation
Arul Mozhi	University of Hyderabad
Asif Iqbal	IIT Patna
Amba Kulkarni	University of Hyderabad
Anil Kumar Singh	IIT BHU, Benaras
Awadhesh Kumar Mishra	CIIL, Mysore
Dafydd Gibbon	Universität Bielefeld, Germany
Daya Krishan Lobiyal	Jawaharlal Nehru University, New Delhi
Dipti Mishra Sharma	IIIT, Hyderabad
Elizabeth Sherley	IITM-Kerala, Trivandrum
Girish Nath Jha	Jawaharlal Nehru University, New Delhi
Hans Uszkoreit	DFKI, Berlin
Indranil Dutta	EFLU, Hyderabad
Jolanta Bachan	Adam Mickiewicz University, Poland
Joseph Mariani	LIMSI-CNRS, France
Jyoti D. Pawar	Goa University
Kalika Bali	MSRI, Bangalore
Karunesh Arora	CDAC Noida
Khalid Choukri	ELRA, France
Lars Hellan	NTNU, Norway
Malhar Kulkarni	IIT Bombay
Manji Bhadra	Bankura University, West Bengal
Massimo Monaglia	University of Florence, Italy
Monojit Choudhary	MSRI Bangalore
Nicoletta Calzolari	ILC-CNR, Pisa, Italy
Niladri Shekhar Dash	ISI Kolkata
Narayan Choudhary	EZDI, Ahmedabad
Panchanan Mohanty	University of Hyderabad
Pushpak Bhattacharya	Director, IIT Patna
Ritesh Kumar	Agra University
R M K Sinha	JSS Academy of Technical Education, Noida
Shivaji Bandhopadhyay	Jadavpur University, Kolkata
Sobha L	AU-KBC Research Centre, Anna University
Soma Paul	IIIT, Hyderabad
S S Aggarwal	KIIT, Gurgaon, India
Subhash Chandra	Delhi University
SwaranLata, Head	TDIL, MCIT, Govt. of India
Umamaheshwar Rao	University of Hyderabad
Vishal Goyal	Punjabi University Patiala
Zygmunt Vetulani	Adam Mickiewicz University, Poland

Introduction

WILDRE – the 3rd Workshop on Indian Language Data: Resources and Evaluation is being organized in Portorož, Slovenia on 24thMay, 2016 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 3rdWILDRE will be

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

The call for papers received a good response from the Indian language technology community. Out of 39 full papers received for review, we selected 7 papers for oral, 5 for short oral, 8 for poster and 4 for demo presentation.

Paper Session I (Oral and Short Oral Presentation)

Tuesday 24 May, 15:00 – 16:00 hrs
Chairperson: **Kalika Bali**

The Festvox Indic Frontend for Grapheme to Phoneme Conversion

Alok Parlikar, Sunayana Sitaram, Andrew Wilkinson and Alan W Black

Text-to-Speech (TTS) systems convert text into phonetic pronunciations which are then processed by Acoustic Models. TTS frontends typically include text processing, lexical lookup and Grapheme-to-Phoneme (g2p) conversion stages. This paper describes the design and implementation of the Indic frontend, which provides explicit support for many major Indian languages, along with a unified framework with easy extensibility for other Indian languages. The Indic frontend handles many phenomena common to Indian languages such as schwa deletion, contextual nasalization, and voicing. It also handles multi-script synthesis between various Indian-language scripts and English. We describe experiments comparing the quality of TTS systems built using the Indic frontend to grapheme-based systems.

Automatic Phonetic Alignment Tool Based on Hidden Markov Model as a Plug-in Tool of Praat for the Languages of Northeast India

Bornali Phukon, Biswajit Dev Sarma, Shakuntala Mahanta and S R M Prasanna

In this paper we present the details of an automatic phonetic alignment method based on a Hidden Markov model method which has been developed as a plug-in tool of the software Praat. This tool has been developed to serve the research interests of the languages of the Northeast India for the first time which are under described and under resourced in many aspects. At present, three Tibeto-Burman languages of Northeast India namely, Tiwa, Dimasa and Kokborok are taken into consideration for the development of this tool, while work on some other languages of the region are under progress. We have collected original speech databases and also prepared phone level transcription for all the three languages. In this phonetic alignment tool, we build HMM models for each phone. Applying the HMM approach, a method of forced alignment is generated whereby phone boundaries are obtained. The tool constitutes primarily of Praat scripts which can be executed from Praat by adding a plug-in. To use the tool, the only requirement is a speech audio file and its corresponding phonemes as input. The outcome depicts tier-wise sentence, word and phonetic alignment of the corresponding spectrograms.

Text Normalization in Urdu Text-to-Speech Synthesis System

Tehreem Waseem, Noorulain Ashraf, Shireen Gull and Sadaf Abdul Rauf

Speech synthesis in Urdu language using Natural Language Processing (NLP) and its development has been researcher’s interest for past three decades. Many major developments in this area have been made recently. Natural Language Processing (NLP) is essential for Text to Speech System (TTS), this process consists of three steps namely: “Text Normalization”, “Text Annotation” and “Phonological Annotation. This paper focuses on the text normalization techniques for TTS system in Urdu Language and elaborates the effects of techniques on the produced speech by the Urdu TTS system.

Classification and Identification of Reduplicated Multi-Word Expressions in Hindi

Renu Singh, Atul Kr. Ojha and Girish Nath Jha

Disambiguating Multi-Word Expressions (MWEs) is often a critical task in NLP applications. Reduplications are an important subclass of MWEs and they are a high-frequency occurrence compared to other kinds of MWEs in Hindi. There are some linguistic challenges in classification and identification of Reduplicated Multiword Expressions (RMWEs) in Hindi. The aim of this paper is to demonstrate linguistic issues pertaining to the distribution of RMWEs, their formalization aspects using a CRF based CRF++ tool and testing and evaluation of the trained system. As per our knowledge, there is no available guideline for annotation of MWEs in Hindi. Therefore, we are presenting the first detailed guidelines for annotation of MWEs in Hindi and it can be applicable in other Indian Languages as well.

Semantic Representation of Tamil Texts using Conceptual Graphs

Pattabhi RK Rao and Sobha Lalitha Devi

This paper presents our work on semantic representation of Tamil documents from sources such as Newswires, Wikipedia articles using conceptual graphs. A conceptual graph is a graph representation for logic based on the semantic networks of artificial intelligence and the existential graphs of Charles Sanders Peirce. A conceptual graph (CG) is a kind of semantic network, which is a network of concept nodes and relation nodes. The challenge in automatic representation of a natural language text in CG is the identification of the concepts and the relationships between them. A concept is an abstract idea conceived mentally by a person. The words of the language are not the concepts but they are the symbols or signs of the concept. Tamil is a Dravidian language. It is a morphologically rich language. In this work we pre-process the text for morph information, Part-of-speech and NP-VP chunks. After preprocessing, the concepts and the kind of relationships between the concepts are identified and the CG is generated. Thus obtained CGs can be further used in applications such as machine translation, information retrieval. In the literature we find that though CGs have been used for English texts and applied for improving the performance of question answering systems, information retrieval systems, but the CG extraction has not been automatic. Semantic representation using CGs for Indian languages is one of the first attempts.

Valance Annotation of Hindi on Typecraft

Srishti Singh

The present paper describes the suitability of TypeCraft framework through valance annotation and construction for Hindi. This paper deals with the different levels of annotation provided on TypeCraft and source for initiating construction labelling using syntactic and semantic information embedded in the language. The annotation challenges in presentation of some major constructions in Hindi like idiomatic expressions, reduplication, conjunct verbs and explicator verbs are discussed along with the construction labelling for Hindi which is a new technique for closer syntactic analysis. This platform also supports more than one free translations and discourse sense labelling for each sentence.

Evaluation of SVM-based Automatic Parts of Speech Tagger for Odia

Pitambar Behera

The authors present an SVM-based POS tagger for Odia language in the paper. The tagger has been trained and tested with Indian Languages Corpora Initiative (ILCI) data of 2, 36, 793 and 1, 28, 646tokens respectively which has been annotated following Bureau of Indian Standards (BIS) annotation scheme. The evaluation has been undertaken under two sections: the statistical and the human, guided by the two approaches of research: quantitative and qualitative. Evaluation results on precision, recall and F measure metrics demonstrate accuracy rates of 93.99%, 92.9971 and 93.49% respectively. So far as the human evaluation is concerned, the agreements are 93.89% (percentage agreement) and 0.87 (Fleiss' Kappa). Finally, the issues and challenges have been discussed in relation to manual annotation and statistical tagger-related issues with a linguistic analysis of errors. On the basis of evaluation results, it can be stated that the present POS tagger is more efficient than the earlier Odia Neural Network tagger (81%) and the SVM tagger (82%) in terms of both accuracy and reliability of the tagger output data.

Poster/Demo Session

Tuesday 24 May, 16:00 – 16:30 hrs
Chairperson: **Massimo Moneglia/Lars Hellan**

Part-of-Speech Tagging System for Maithili Tweets

Divyanshu Bhardwaj and Amitava Das

Part-of-Speech (POS) tagging is the prerequisite for all Natural Language Processing (NLP) applications be it Sentiment Analysis, Natural Language Parsing, Word Sense Disambiguation, Text to Speech Processing or Information Retrieval among others. Maithili is one of the staple languages of Bihar. It is spoken by approximately 34.7 million people as of 2000. Despite the fact that POS tagging for major Indian languages has been done in recent years, Maithili has not been delved into much. Consequently, developing a Part-of-Speech (POS) tagging system for Maithili is vital. This paper deals with the collection and developing an automated system for POS tagging at word level for Maithili Tweets using both coarse-grained and fine-grained tagsets. Although code-mixing with English, that too, Indian languages written in romanized phonetics is the prevalent practice in Indian social media but in this paper, only monolingual tweets, written in Devanagari are considered. The efficiency of different tagging systems based upon four machine learning algorithms (Naive Bayes, Sequential Minimal Optimization, Random Forest and Conditional Random Field) is noted.

Classification and Resolution of Linguistic Divergences in English-Urdu Machine Translation

Sharmin Muzaffar, Pitambar Behera and Girish Nath Jha

The present paper attempts at exploring, classifying and resolving various types of divergence patterns in the context of English-Urdu Machine Translation where English and Urdu are SL and TL respectively. So far as the methodology is concerned, we have observed the English-Urdu pair sentences and analyzed the translated output taking into consideration different areas of translational divergences. We have taken one thousand corpus of the ILCI English sentences for this study and analyzed the translated Urdu output in bulk taking into consideration different areas of translational divergences on web-based Machine Translation platforms namely, Bing and Google Translate. Dorr's (1994, 1995) theoretical framework has been adopted for the classification and resolution of the linguistic divergences in this undertaken study. Dorr has classified the divergences into two broad categories (lexical-semantic and syntactic divergences) and proposed Lexical Conceptual Structure-based resolution for them. This study would help identify, classify and resolve the underlying divergence patterns between these languages so as to develop MT systems considering the divergence errors and enhance the performance of the MT.

Extractive based Email Summarization: An Unsupervised Hybrid approach using Graph Based Sentence Ranking and K-Means Clustering algorithm

Milind Shivolkar, Jyoti D. Pawar and S. Baskar

Over the years, Automatic Text Summarization is widely studied by many researchers. Here, an attempt is made to generate an automatic summary of a given text document based on an unsupervised hybrid model. The model comprises of an extractive method: a Graph-based text ranking and K-means: a clustering algorithm. Ranked sentences are obtained using the graph-theoretic ranking model here word frequency, word position, and string pattern based ranking are calculated. The K-Means algorithm generates the coherent topic clusters. Using the output of Graph-based method and K-means clusters, Sentence Importance Score(SIS) is calculated for each sentence, where top 70% ranked sentences and centralised topics of each cluster (excluding those topics which fall in the outlier zone) are used. The unsupervised hybrid approach is an attempt to inherit one of the human practice of reading and then summarizing the text in short while keeping the original insight of that text by the virtue of important sentences and keywords. The system is tested on dataset for Summarization and Keyword Extraction from Emails which on evaluation gives an average of 0.57 score on ROUGE 2.0 tool.

Konkani SentiWordNet-Resource For Sentiment Analysis Using Supervised Learning Approach

Ashweta Fondekar, Jyoti D. Pawar and Ramdas Karmali

Sentiment Analysis (SA) is the process of analyzing and predicting the hidden attitude/opinion in the given text expressed by an individual. Till now, ample amount of work has been carried out for the English language. But, no work is performed for the language Konkani in the field of Sentiment Analysis. Lexicon-based SA is a good beginning for any language, especially if the digital content is limited. Hence, the main motive of this paper is; to present the sentiment lexicon called SentiWordNet for Konkani language. The process of creating Konkani SentiWordNet is under progress using the Supervised Learning Approach. In this approach, the training set is generated using a Synset Projection Approach and Support Vector Machine (SVM) algorithm to classify the data. The reason behind using the Synset Projection Approach for building a training dataset is; English Sentiwordnet is developed using Semi-Supervised Approach where the training dataset is generated using WordNet lexical relations but; in Konkani WordNet, lexical relations are not yet developed. Hence, Synset Projection Approach is preferred. Conducted experimental results for the proposed algorithm are reported in this paper.

Verb Mapping: A Dilemma in Sanskrit-Hindi Machine Translation

Kumar Nripendra Pathak and Girish Nath Jha

Creating a Fully Automated Machine Translation is a challenge. MT system developers have to take care of all minute aspects of both the language pairs (i.e. the Source Language and the Target Language). Issue of verb mapping between language pairs needs a careful study of verb pattern of those languages. A close look towards verb pattern indicates the importance of the conditional use of verb forms in a language. The conditional use of the verbs is a challenge for MT systems. As Sanskrit-Hindi Machine Translation (SHMT) is an ongoing task and the Sanskrit Consortium, funded by the DIT, Govt. of India, has already finished its first Phase, this study becomes more relevant. The proposed SHMT–Sampark System is not functional yet. An Interface of SHMT-Anusaaraka is available on the website of Sanskrit Department, HCU, Hyderabad. In this study, some challenging aspects of verb mapping have been noticed as the dilemma in SHMT.

Lexical Resources for Recognition, Analysis and Word Formation process for Sanskrit Morphology

Subhash Chandra, Bhupendra Kumar, Vivek Kumar and Sakshi

Aṣṭādhyāyī (AD) which particularly contains about 3,959 rules of Sanskrit morphology, syntax and semantics. Sanskrit word formation process is taught in all major Indian Universities offering Sanskrit courses at Undergraduate (UG) and post graduate (PG) level. This paper introduces a web based word formation process tools for students and teachers of Sanskrit with the aim of teaching and learning Sanskrit morphological inflectional process based on Pāṇini rules and *prakriyāgranthas* of AD. The system is developed by combining rule and example based approaches used by Pāṇini. There are three components entitled Recognizer, Analyzer and Word Formation Process (WFP) Generator in the system that generate word formation process for *subanta* (nominal), primary verb (*tiṅanta*) and secondary verb (*sanādyanta*). Currently this system covers *subanta* and *tiṅanta* only and it is being used by the Sanskrit students and teachers for learning and teaching Sanskrit Grammar.

Building a Statistical tagger for Sanskrit

Archana Tiwari

Abstract In this paper, the author is discussing on part of speech tagging of Sanskrit and the development of a tagger using Support Vector Machine. The Data for the training of machine has been taken from literature and general domain. Data has been taken from ‘Panchatantra’ and ‘Sudharama’ Sanskrit newspaper and various blogs. In this process of data tagging the BIS tagset for Sanskrit, has been used. The system has adopted the statistical learning approach. In which the system will learn from annotated corpus and apply it on unseen text. The tagger has achieved 82% and 80.89% accuracy till now. The paper is divided in three parts. In the first part, the part of speech tagging and its importance and development of corpus and tagging methods have been explored. In the second part the development of tagger, its training, evaluation and result are discussed. In the third part the issues and challenge has been analysed.

Developing annotated multimodal corpus for automatic recognition of verbal aggression in Hindi

Ritesh Kumar, Atul Kr. Ojha and Bornini Lahiri

Verbal aggression could be defined as any act which seeks to disturb the social and relational equilibrium. In a large number of cases, verbal aggression could be a precursor to certain kind of criminal activities; in others, as in political speeches, it might be desirable to take note of specific kinds of aggression. In this paper we discuss the development of a multimodal corpus of Hindi which could be used for automatically recognising verbal aggression in Hindi. The complete raw corpus currently consists of approximately 1000 hours of audio-visual debates and discussions carried out in the Indian Parliament (and made available for research) as well as some recordings from different news channels available in public domain over the web. Out of this, approximately 30 hours have already been transcribed and around 5 hours is annotated using the aggression tagset. In this paper, we also discuss this tagset which is being used for annotation.

Demo: SwiftKey Keyboard for Indian Languages

Jalpa Zaladi and Caroline Gasperin

This paper describes a text input method for touch screen typing in Indian languages on mobile phones and tablets. Typing in native script on mobile is quite difficult due to the complicated structure of Indic scripts. Moreover, limited mobile screen space restricts the number of characters and symbols displayed on the screen. People typing in native script face difficulties in learning various layouts and in finding required letters among a relatively large set of consonants, vowels and conjunct consonants. SwiftKey app provides a solution that makes typing in native script easier and enjoyable by carefully packaging a text prediction system with dynamic language-specific layouts. It also supports features such as emoji prediction, gesture typing and themes. SwiftKey supports all 22 official Indian languages.

Translating Code-Mixed Tweets: A Language Detection Based System

Shruti Rijhwani, Royal Sequeira, Monojit Choudhury and Kalika Bali

We demonstrate a system for the machine translation of code-mixed text in several Indian and European languages. We first perform word-level language detection and matrix language identification. We then use this information and an existing translator in order to translate code-mixed tweets into a language of the user's choice.

Demo Proposal: Tamil-Hindi Automatic Machine Translation – A detailed Description

Sobha Lalitha Devi, Vijay Sundar Ram, Sindhuja Gopalan, Pattabhi RK Rao and Lakshmi S

Automatic machine translation between two languages, poses various challenges. We have presented a detailed description on machine translation between Tamil and Hindi, where Tamil belongs to Dravidian language family and Hindi belongs to Indo-Aryan language family. These two languages have similarities such as verb final, morphological richness, relatively free-word order and they are structurally dissimilar. We have described the methodology and challenges addressed in each module.

A Demo Proposal: Tamil-English Cross Lingual Information Access (CLIA) System

Sobha Lalitha Devi, Pattabhi RK Rao, Vijay Sundar Ram and C.S Malarkodi

In this paper we present Tamil-English Cross Lingual Information Access (CLIA) system. The objective of this system is to provide users who are non–English speakers, to access information available on internet in their own native language, Tamil. This system enables users to give queries in Tamil and retrieve documents in Tamil as well as in English. The user given query is translated to English for retrieving English documents. Query translation is done using synset dictionaries, bilingual dictionaries and transliteration. The content of English documents is translated to Tamil using Template translation. The main modules of the system are i) Crawling ii) Input Processing iii) Indexing iv) Query Processing v) Ranking vi) Output Processing. We have obtained a MAP score of 0.3980 for the Tamil – English cross lingual search. The results are encouraging and comparable with the state of the art. The system is deployed and accessible through web link.

Paper Session II (Oral Presentation)

Tuesday 24 May, 16:30 – 17:30 hrs
Chairperson: **Zygmunt Vetulani**

Towards a Big Data View on South Asian Linguistic Diversity

Lars Borin, Shafqat Mumtaz Virk and Anju Saxena

South Asia with its rich and diverse linguistic tapestry of hundreds of languages, including many from four major language families, and a long history of intensive language contact, provides rich empirical data for studies of linguistic genealogy, linguistic typology, and language contact. South Asia is often referred to as a linguistic area, a region where, due to close contact and widespread multilingualism, languages have influenced one another to the extent that both related and unrelated languages are more similar on many linguistic levels than we would expect. However, with some rare exceptions, most studies are largely impressionistic, drawing examples from a few languages. In this paper we present our ongoing work aiming at turning the linguistic material available in Grierson's Linguistic Survey of India (LSI) into a digital language resource, a database suitable for a broad array of linguistic investigations of the languages of South Asia. In addition to this, we aim to contribute to the methodological development of large-scale comparative linguistics drawing on digital language resources, by exploring NLP techniques for extracting linguistic information from free-text language descriptions of the kind found in the LSI.

A Hybrid Chunker for Hindi and Indian English

Atul Kr. Ojha, Srishti Singh, Pitambar Behera and Girish Nath Jha

The paper presents a CRF based hybridized chunker for Hindi and Indian English. The immediate goal is to chunk text data in the ILCI project funded by DeitY, Govt of India. The experiment was conducted on 25k annotated sentences on the data from health and tourism domains. 23k sentences were used for training and the rest 2k sentences were used for evaluation. The experiment involved the following stages: training the chunker, automatic chunking and validation of chunked output for Hindi and Indian English; and finding measures to solve issues detected at different levels of experiment. The chunker for Indian English is developed on ILMT chunk tag scheme to meet the necessary mapping requirements of the translation tool for English to Indian languages. The accuracies of Hindi and Indian English chunker are 88.84% & 89.04 %, respectively. So far as Hindi chunker is concerned, we have observed errors in the chunk categories such as noun (pronominal), verb finite, verb non-finite (conjunct verb), adjectival phrase etc. Errors like finite-non-finite, adverb-conjunction, wh-determiner and conjunction chunk etc are discussed in detail for the development of English chunker. Implementation of hybrid approach for error resolution has also been attempted.

Machine Translating Code Mixed Text: Pain Points and Sweet Spots

Akshay Gadre, Rafiya Begum, Kalika Bali and Monojit Choudhury

In this qualitative evaluation of a state-of-the-art SMT system, we study the performance on Hindi-English code-mixed tweets. Our study indicates that (a) language identification and transliteration can go a long way in improving the performance, (b) translation to the matrix language gives better results, and (c) quality of translation heavily depends on the number of switch-points and the nature of the embedded linguistic unit(s).

Discovering Thematic Knowledge from Code-Mixed Chat Messages Using Topic Model

Kavita Asnani and Jyoti D. Pawar

In current times, the trend of mixing two or more languages together (code-mixing) in communication on social media is very popular. Such code-mixed chat data is enormously generated and is usually noisy, sparse and exhibits high dispersion of useful topics which people discuss. In such a scenario, it is very challenging to automatically extract relevant thematic information which contributes to useful knowledge. In order to discover latent themes from multilingual data, a standard topic model called Probabilistic Latent Semantic Analysis (PLSA) is used in existing literature. However, it addresses the inter-sentence multilingualism. In this paper, we propose a novel method which is basically based on co-occurrences of words within a code-mixed message. Thus built co-occurrence matrix for chat is exposed to PLSA which is used to discover thematic knowledge from it. In such code-mixed chat text, inter-sentence, intra-sentence and intra-word level code mixing may randomly occur. We have proved with extensive experiments that it is possible to use this strategy to discover latent themes from semantic topic clusters. We tested our system using FIRE 2014 dataset.

Evaluation of Anuvadaksh (EILMT) English-Odia Machine-assisted Translation Tool

Pitambar Behera, Renu Singh and Girish Nath Jha

The authors present the evaluation of the Anuvadaksh English to Odia Machine Translation System which has been developed by the Applied Artificial Intelligence Group (AAI-G) of C-DAC, Pune applying the MANTRA Technology from English to eight Indian Languages in EILMT (English-Indian Languages Machine Translation) Consortium, under the TDIL (Technology Development for Indian Languages), Deity (Dept. of Electronics and Information Technology), Government of India. In this study, a 1k ILCI English sentence corpus has been used from the domain of health as input to evaluate the web-based system output in Odia. For evaluating the output qualitatively, the Inter-translator Agreement of three human evaluators with scores on the five point scale has been taken into consideration. The scores have been calculated by the Fleiss’ Kappa statistics in terms of reliability and adequacy on the basis of which linguistic error analysis and suggestions for improvement have been provided. The Kappa scores for reliability and adequacy are 0.30 and 0.28 respectively which refer to the fact that both are fair.

ETHI-CA^2 2016:
ETHics In Corpus Collection, Annotation & Application

24 May 2016

ABSTRACTS

Editors:

**Laurence Devillers, Björn Schuller, Emily Mower Provost, Peter Robinson,
Joseph Mariani, Agnes Delaborde**

Workshop Programme

09:00 – Introduction
Laurence Devillers

09:10 – Keynote
Chairperson: Laurence Devillers
Edouard Geoffrois, *Interactive System Adaptation: Foreseen Impacts on the Organisation and Ethics of System Development*

09:50 – Talk 1
Chairperson: Björn Schuller
Teresa Scantamburlo and Marcello Pelillo, *Contextualizing Privacy in the Context of Data Science*

10:10 – Talk 2
Chairperson: Björn Schuller
Kevin Bretonnel Cohen, Karen Fort, Gilles Adda, Sophia Zhou and Dimeji Farri, *Ethical Issues in Corpus Linguistics And Annotation: Pay Per Hit Does Not Affect Effective Hourly Rate For Linguistic Resource Development On Amazon Mechanical Turk*

10:30 – Coffee and Poster session
Chairperson: Laurence Devillers

- Jana Diesner and Chieh-Li Chin, *Gratis, Libre, or Something Else? Regulations and Misassumptions Related to Working with Publicly Available Text Data*
- Wessel Reijers, Eva Vanmassenhove, David Lewis and Joss Moorkens, *On the Need for a Global Declaration of Ethical Principles for Experimentation with Personal Data*
- Agnes Delaborde and Laurence Devillers, *Diffusion of Memory Footprints for an Ethical Human-Robot Interaction System*
- Björn Schuller, Jean-Gabriel Ganascia and Laurence Devillers, *Multimodal Sentiment Analysis in the Wild: Ethical considerations on Data Collection, Annotation, and Exploitation*
- Jocelynn Cu, Merlin Teodosia Suarez and Madelene Sta. Maria, *Subscribing to the Belmont Report: The Case of Creating Emotion Corpora*
- Lucile Béchade, Agnes Delaborde, Guillaume Dubuisson Duplessis and Laurence Devillers, *Ethical Considerations and Feedback from Social Human-Robot Interaction with Elderly People*

11:10 – Talk 3
Chairperson: Joseph Mariani
Joss Moorkens, David Lewis, Wessel Reijers and Eva Vanmassenhove, *Language Resources and Translator Disempowerment*

11:30 – Talk 4
Chairperson: Joseph Mariani
Simone Hantke, Anton Batliner and Björn Schuller, *Ethics for Crowdsourced Corpus Collection, Data Annotation and its Application in the Web-based Game iHEARu-PLAY*

11:50 – Talk 5
Chairperson: Joseph Mariani
Agnes Delaborde, Noémie Enser, Alexandra Bensamoun and Laurence Devillers, *Liability Specification in Robotics: Ethical and Legal Transversal Regards*

12:10 – Panel and open discussion
Chairperson: Björn Schuller

12:30 – Conclusion
Laurence Devillers

Workshop Organizers

Laurence Devillers	LIMSI-CNRS/Paris-Sorbonne University, France
Björn Schuller	Imperial College London, UK/University of Passau, Germany
Emily Mower Provost	University of Michigan, USA
Peter Robinson	University Cambridge, UK
Joseph Mariani	IMMI/LIMSI-CNRS/Paris-Saclay University, France
Agnes Delaborde	LIMSI-CNRS/CERDI/Paris-Saclay University, France

Workshop Programme Committee

Gilles Adda	LIMSI-CNRS, France
Jean-Yves Antoine	University of Tours, France
Nick Campbell	TCD, Ireland
Alain Couillault	GFII, France
Anna Esposito	UNINA, Italy
Karën Fort	Université Paris-Sorbonne, France
Jean-Gabriel Ganascia	UPMC, France
Alexei Grinbaum	CEA, France
Hatice Gunes	Queen Mary University of London, UK
Dirk Heylen	University of Twente, Netherlands
Catherine Tessier	ONERA, France
Isabelle Trancoso	INESC, Portugal
Guillaume Dubuisson Duplessis	LIMSI-CNRS, France

Preface

Description

The focus of ETHI-CA^2 spans ethical aspects around the entire processing pipeline from speech and language, as well as multimodal resource collection and annotation, to system development and application. In the recent time of ever-more collection "in the wild" of individual and personal multimodal and multi-sensorial "Big Data", crowd-sourced annotation by large groups of individuals with often unknown reliability and high subjectivity, and "deep" and autonomous learning with limited transparency of what is being learnt, and how applications such as in health or robotics depending on such data may behave, ethics have become more crucial than ever in the field of language and multimodal resources. This makes ethics a key concern of the LREC community. There is, however, a surprising if not shocking white spot in the landscape of workshops, special session, or journal special issues in this field, which ETHI-CA^2 aims to fill in.

The goal is thus to connect individuals ranging across LREC's fields of interest such as human-machine and robot- and computer-mediated human-human interaction and communication, affective, behavioral, and social computing whose work touches on crucial ethical issues (e.g. privacy, tracability, explainability, evaluation, responsibility, etc.). According systems increasingly interact with and exploit data from humans of all ranges (e.g. children, adults, vulnerable populations) including non-verbal and verbal data occurring in a variety of real-life contexts (e.g. at home, the hospital, on the phone, in the car, classroom, or public transportation) and act as assistive and partially instructive technologies, companions, and/or commercial or even decision-making systems. Obviously, an immense responsibility lies at the different ends from data recording, labeling, and storage, to its processing and usage.

Motivation

Emerging interactive systems have changed the way we connect with our machines, modifying how we socialize, our reasoning capabilities, and our behavior. These areas inspire critical questions centering on the ethics, the goals, and the deployment of innovative products that can change our lives and society. Many current systems operate on private user data, including identifiable information, or data that provides insight into an individual's life routine. The workshop will provide discussions about user consent and the notion of informed data collection.

Cloud-based storage systems have grown in popularity as the scope of user-content and user-generated content has greatly increased in size. The workshop will provide discussions on best practices for data annotation and storage and evolving views on data ownership.

Systems have become increasingly capable of mimicking human behavior through research in affective computing. These systems have provided demonstrated utility, for interactions with vulnerable populations (e.g. the elderly, children with autism). The workshop will provide discussions on considerations for vulnerable populations.

The common mantra for assistive technology is,"augmenting human care, rather than replacing human care". It is critical that the community anticipates this shift and understands the implication of machine-in-the-loop diagnostic and assessment strategies.

Topics of interest

Topics include, but are not limited to:

- Ethics in recording of private content
- Ethics in multimodal, sensorial data collection
- Ethics in annotation (crowd-sourced) of private data
- Data storage/sharing/anonymization
- Transparency in Machine Learning
- Ethics in Affective, Behavioural, and Social Computing
- Responsibility in Educational Software and Serious Games
- Human-machine interaction for vulnerable populations
- Computer-mediated Human-Human Communication
- Responsibility in Decision-Support based on Data
- The role of assistive technology in health care

Summary of the call

The ETHI-CA^2 2016 workshop is a crucially needed first edition in a planned for longer series. The goal of the workshop is to connect individuals ranging across LREC's fields of interest such as human-machine and robot- and computer-mediated human-human interaction and communication, affective, behavioural, and social computing whose work touches on crucial ethical issues (e.g. privacy, traceability, explainability, evaluation, responsibility, etc.). These areas inspire critical questions centering on the ethics, the goals, and the deployment of innovative products that can change our lives and consequently, society. It is critical that our notion of ethical principles evolves with the design of technology. As humans put increasing trust in systems, we must understand how best to protect privacy, explain what information the systems record, the implications of these recordings, what a system can learn about a user, what a third party could learn by gaining access to the data, changes in human behavior resulting from the presence of the system, and many other factors. It is important that technologists and ethicists maintain a conversation over the development and deployment lifecycles of the technology. The ambition of this workshop is to collect the main ethics, goals and societal impact questions of our community including experts in sociology, psychology, neuroscience or philosophy. At LREC 2016, the workshop shall encourage a broad range of its community's researchers to reflect about and exchange on ethical issues inherent in their research, providing an environment in which ethics co-evolve with technology.

Keynote

Edouard Geoffrois

Edouard Geoffrois did his PhD in the field of automatic speech recognition, after graduating from the Ecole Polytechnique and getting a Master Degree in cognitive science. He joined the French defense procurement agency (DGA) in 1996, where he created the activities in speech and language processing. He has initiated and managed several projects and programs in the field of multimedia information processing, including evaluation campaigns, and has elaborated new metrics and evaluation protocols when needed to steer the developments. Since 2015, he is on loan to the French national research agency (ANR), where he coordinates two international ERA-NET programs (CHIST-ERA and FLAG-ERA).

Interactive System Adaptation: Foreseen Impacts on the Organisation and Ethics of System Development

Interactive System Adaptation is the ability of a system to learn from user feedback. It can be seen as a special case of autonomous learning, which is the ability of a system to learn from a new environment, without any intervention from its initial developers, and to increase its performance in this new environment while maintaining it in the initial one. While system adaptation is an active topic of research, interactive system adaptation and autonomous learning have never been evaluated in a comparative way, and much research remains needed to develop such capabilities. Following the proposal of new evaluation protocols for interactive system adaptation, such developments can be expected to happen at a much faster pace in the coming years. These capabilities, when available, will have important implications on the organisation and ethics of system development. The most obvious one is that the user's data will not have to be provided to the initial developer to ensure the adaptation, and will remain under the control of the user instead. Another one is that the behavior of the systems will be influenced by several types of actors. The objective evaluation of such systems will therefore become important not only to measure their technical capabilities but also to attribute the responsibilities of these actors in case of inadequate system behavior.

Talk 1

09:50 – 10:10
Chairperson: Björn Schuller

Contextualizing Privacy in the Context of Data Science

Teresa Scantamburlo, Marcello Pelillo

Privacy is one of the most long-standing social issues associated to the development of information and communication technology and, over the years, from the diffusion of large databases to the rise of the World Wide Web and today's Internet of Things, just to name a few examples, the concerns have been intensified with consequences on the public discussion, design practices and policy making. Unfortunately the plethora of technical details on this topic has often discouraged people from tackling the problem of privacy and, hence, from being really active in the discussion of proposed approaches and solutions. In this paper we would like to provide fresh motivations to the inclusion of privacy in the overall pipeline of data processing, making this notion and its related issues somewhat more accessible to a non-expert audience. We will do that not by promoting new design standards or methodologies, which would add further technicalities, but by developing a critical perspective on the current approaches. In this way, we aim at providing data specialists with novel conceptual frameworks (e.g. the theory of conceptual integrity) to evaluate and better understand the place of privacy in their own work.

Talk 2

10:10 – 10:30
Chairperson: Björn Schuller

Ethical Issues in Corpus Linguistics And Annotation: Pay Per Hit Does Not Affect Effective Hourly Rate For Linguistic Resource Development On Amazon Mechanical Turk

Kevin Bretonnel Cohen, Karen Fort, Gilles Adda, Sophia Zhou and Dimeji Farri

Ethical issues reported with paid crowdsourcing include unfairly low wages. It is assumed that such issues are under the control of the task requester. Can one control the amount that a worker earns by controlling the amount that one pays? 412 linguistic data development tasks were submitted to Amazon Mechanical Turk. The pay per HIT was manipulated through a range of values. We examined the relationship between the pay that is offered per HIT and the effective pay rate. There is no such relationship. Paying more per HIT does not cause workers to earn more: the higher the pay per HIT, the more time workers spend on them (R = 0:92). So, the effective hourly rate stays roughly the same. The finding has clear implications for language resource builders who want to behave ethically: other means must be found in order to compensate workers fairly. The findings of this paper should not be taken as an endorsement of unfairly low pay rates for crowdsourcing workers. Rather, the intention is to point out that additional measures, such as pre-calculating and communicating to the workers an average hourly, rather than per-task, rate must be found in order to ensure an ethical rate of pay.

Coffee and Poster session

10:30 – 11:10

Chairperson: Laurence Devillers

Gratis, Libre, or Something Else? Regulations and Misassumptions Related to Working with Publicly Available Text Data

Jana Diesner and Chieh-Li Chin

Raw, marked up, and annotated language resources have enabled significant progress with science and applications. Continuing to innovate requires access to user generated and professionally produced, publicly available content, such as data from online production communities, social networking platforms, customer review sites, discussion forums, and expert blogs. However, researchers do not always have a comprehensive or correct understanding of what types of online data are permitted to be collected and used in what ways. This paper aims to clarify this point. The way in which a dataset is "open" is not defined by its accessibility, but by its copyright agreement, license, and possibly other regulations. In other words, the fact that a dataset is visible free of charge and without logging in to a service does not necessarily mean that the data can also be collected, analyzed, modified, or redistributed. The open software movement had introduced the distinction between free as in "free speech" (freedom from restriction, "libre") versus free as in "free beer" (freedom from cost, "gratis"). A possible risk or misassumption related to working with publicly available text data is to mistake gratis data for libre when some online content is really just free to look at. We summarize approaches to responsible and rule-compliant research with respect to "open data".

On the Need for a Global Declaration of Ethical Principles for Experimentation with Personal Data

Wessel Reijers, Eva Vanmassenhove, David Lewis and Joss Moorkens

In this paper, we argue that there is a growing need for a globally accepted set of ethical principles for experimentation that makes use of collections of personal multimodal data. Just as the Helsinki declaration was signed in 1964 to provide ethical principles that would guide all experimentation with human subjects, we argue that today the "digital personae" ought to be protected by a similar, globally endorsed declaration that informs legal regulations and policy making. The rationale for such a declaration lies in the increasing pervasiveness of the use of personal data in many aspects of our daily lives, as well as in the scattered nature of data research for which particular implementations of research ethics at the level of a single institution would not suffice. We argue that the asymmetry between ethical standards of public and of commercial entities, the borderless and boundless nature of online experimentation and the increasing ambiguity of the meaning of online "experiments" are compelling reasons to propose a global declaration of ethical principles for experimentation with personal data.

Diffusion of Memory Footprints for an Ethical Human-Robot Interaction System

Agnes Delaborde and Laurence Devillers

Along with the constant amelioration of the artificial intelligence skills in robots, has arisen a strong will among the community to define ethical limits to the behaviors of the robots. The implementation of ethics and morality in an autonomous system represents a research challenge, and several practical propositions have been offered by the community. The authors propose trails for a Human-Robot Interaction architecture in which the selective diffusion of footprints and logs extracted from the robot's memory (low-level inputs, interpretation, decisions, actions) would improve the traceability of the robot's internal decision-making, which could for example offer a guarantee of transparency in case of faulty or contentious situations. The description of the proposed architecture is based on the authors' studies on social Human-Robot Interaction systems designed in the context of the French robotic project ROMEO. The authors' proposition will be subsequently assessed in the course of a French transdisciplinary project involving the fields of robotics, law and artificial intelligence.

Multimodal Sentiment Analysis in the Wild: Ethical considerations on Data Collection, Annotation, and Exploitation

Björn Schuller, Jean-Gabriel Ganascia and Laurence Devillers

Some ethical issues that arise for data collection and annotation of audiovisual and general multimodal sentiment, affect, and emotion data "in the wild" are of types that have been well explored, and there are good reasons to believe that they can be handled in routine ways. They mainly involve two areas, namely research with human participants, and protection of personal data. Some other ethical issues coming with such data such as its exploitation in real-life recognition engines and evaluation in long-term usages are, however, less explored. Here, we aim to discuss both – the more "routine" aspects as well as the white spots in the literature of the field. The discussion will be guided by needs and observations as well as plans made during and for the European SEWA project to provide a showcase example.

Subscribing to the Belmont Report: The Case of Creating Emotion Corpora

Jocelynn Cu, Merlin Teodosia Suarez and Madelene Sta. Maria

Works on human emotion and behavior analysis has been on the increase in recent years. This is primarily the result of maturity of information technology, evidence that it has enmeshed itself in a human's everyday activities. Current approaches to emotion and behavior modeling require the creation of corpora from human subjects typically engaged in interactions. Collection of information and other data from humans necessitates following certain guidelines to ensure their privacy, security and over-all well-being. This work presents how the Belmont Report may be adapted in the practice and review of acceptable standards in the creation of emotion corpora for developing countries, given that these communities do not possess high awareness of their privacy rights. It describes the creation of several multi-modal corpora of patients and students, including the ethical practices employed following the principles indicated in the Belmont Report. At the end of the paper, recommendations how to better improve research practices are shared, including possible research directions.

Ethical Considerations and Feedback from Social Human-Robot Interaction with Elderly People

Lucile Bechade, Agnes Delaborde, Guillaume Dubuisson Duplessis and Laurence Devillers

Field studies in Human-Robot Interaction are essential for the design of socially acceptable robots. This paper describes a data collection carried out with twelve elderly people interacting with the Nao robot via a Wizard-of-Oz system. This interaction involves two scenarios implementable in a social robot as an affective companion in everyday life. One of the scenarios involves humor strategies while the other one involves negotiation strategies. In this paper, the authors detail the designs of the system, the scenarios and the data collection. The authors take a closer look at the opinions of the elderly collected through self-reports and through a private verbal exchange with one experimenter. These opinions include recommendations about the features of the robot (such as speech rate, volume and voice) as well as points of view about the ethical usage of affective robots.

Talk 3

11:10 – 11:30
Chairperson: Joseph Mariani

Language Resources and Translator Disempowerment

Joss Moorkens, David Lewis, Wessel Reijers and Eva Vanmassenhove

Language resources used for machine translation are created by human translators. These translators have legal rights with regard to copyright ownership of translated texts and databases of parallel bilingual texts, but may not be in a position to assert these rights due to employment practices widespread in the translation industry. This paper examines these employment practices in detail, and looks at the legal situation for ownership of translation resources. It also considers the situation from the standpoint of current owners of resources.

Talk 4

11:30 – 11:50
Chairperson: Joseph Mariani

Ethics for Crowdsourced Corpus Collection, Data Annotation and its Application in the Web-based Game iHEARu-PLAY

Simone Hantke, Anton Batliner and Björn Schuller

We address ethical considerations concerning iHEARu-PLAY, a web-based, crowdsourced, multiplayer game for large-scale, real-life corpus collection and multi-label, holistic data annotation for advanced paralinguistic tasks. While playing the game, users are recorded or perform labelling tasks, compete with other players, and are rewarded with scores and different prizes. Players will have fun playing the game and at the same time support science. With this modular, cross-platform crowdsourcing game, different ethical and privacy issues arise. A closer look is taken on ethics in recording of private content, data collection, data annotation, and storage, as well as sharing the data within iHEARu-PLAY. Further, we address the interplay of science and society in ethics and relate this with our application iHEARu-PLAY.

Talk 5

11:50 – 12:10

Chairperson: Joseph Mariani

Liability Specification in Robotics: Ethical and Legal Transversal Regards

Agnes Delaborde, Noémie Enser, Alexandra Bensamoun and Laurence Devillers

Beyond the implementation of moral considerations, an ethical robot should be designed in a way that foresees the potential damages it could cause, and that also anticipates the way the human beings in its environment (from the designer to the user) could be held responsible for its acts. In this present study, the authors offer to consider the actions of the robot under the French liability regime for the actions of things. In these conditions, the designer, the manufacturer and the user could be held liable for the actions of the robot. As a preventive measure, the robot would be subject to a mandatory insurance assumed by the user, and the designer would endow the interface with a log of data that could be assessed by an expert. As part of a curative approach, the authors present realistic case studies in which the robot could cause damage, with a determination of the liability based upon the liability regime for the action of things.

International Workshop on Social Media World Sensors (Sideways)

24 May 2016

ABSTRACTS

Editors:

Mario Cataldi, Luigi Di Caro, Claudio Schifanella

Workshop Programme

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

09:30 – 10:00 – Dane Bell, Daniel Fried, Luwen Huangfu, Mihai Surdeanu, Stephen Kobourov, *Challenges for using social media for early detection of T2DM (invited talk)*

10:00 – 10:30 – Tim Kreutz and Malvina Nissim, *Catching Events in the Twitter Stream: A showcase of student projects*

10:30 – 11:00 Coffee break

11:00 – 11:30 – Udo Krushwitz, Ayman Al Helbawy, Massimo Poesio, *Exploiting Social Media to Address Fundamental Human Rights Issues (invited talk)*

11:30 – 12:00 – Emanuele Di Rosa, Alberto Durante, *App2Check: a Machine Learning-based system for Sentiment Analysis of App Reviews in Italian Language*

12:00 – 12:30 – Christian Colella, *Distrusting Science on Communication Platforms: Socio-anthropological Aspects of the Science-Society Dialectic within a Phytosanitary Emergency*

12:30 – 13:00 – Luca Vignaroli, Claudio Schifanella, K. Selcuk Candan, Ruggero Pensa, Maria Luisa Sapino, *Tracking and analyzing the "second life" of TV content: a media and social-driven framework (invited industrial demo)*

Workshop Organizers

Mario Cataldi	Université Paris 8, France
Luigi Di Caro	Department of Computer Science – University of Turin, Italy
Claudio Schifanella	RAI – Centre for Research and Technological Innovation, Turin, Italy

Workshop Programme Committee

Luca Aiello	Yahoo! Labs, London
Andrea Ballatore	University of California, Santa Barbara, USA
Iván Cantador	Universidad Autónoma de Madrid, Spain
Federica Cena	University of Torino, Italy
Martin Chorley	Cardiff University, Wales
Emilio Ferrara	Indiana University Bloomington, USA
Simon Harper	University of Manchester, England
Dino Ienco	Irstea, UMR TETIS, Montpellier, France
Séamus Lawless	Trinity College Dublin, Ireland
Emmanuel Malherbe	Multiposting, France
Rosa Meo	University of Torino, Italy
Ruggero G. Pensa	University of Torino, Italy
Rossano Schifanella	University of Torino, Italy
Thomas Steiner	Google, USA
Luca Vignaroli	RAI – Centre for Research and Technological Innovation, Turin, Italy

Introduction

It is our great pleasure to welcome you to the *2016 ACM Workshop on Social Media World Sensors - Sideways 2016,* which is held in conjunction with the *10th Edition of its Language Resources and Evaluation Conference.- LREC 2016* in Portoroz, Slovenia.

This second edition of the workshop aims at bringing together academics and practitioners from different areas to promote the vision of social media as *social sensors*. Nowadays, social platforms have become the most popular communication system all over the world. In fact, due to the short format of messages and the accessibility of these systems, users tend to shift from traditional communication tools (such as blogs, web sites and mailing lists) to social network for various purposes. Billions of messages are appearing daily in these services such as Twitter, Tumblr, Facebook, etc. The authors of these messages share content about their private life, exchanging opinions on a variety of topics and discussing a wide range of information news.

Even if this system cannot represent an alternative to the authoritative information media, considering the number of its users and the impressive response time of their contributions, they represent a sort of real-time news sensor that can also predate the best newspapers in informing the web community about the emerging topics and trends. In fact, the most important information media always need a certain amount of time to react to a news event; i.e. professional journalists require time, collaborators and/or technology support to provide a professional report. However, a user can easily report, in few characters, what is happening in front of the user's eyes, without any concern about the readers or the writing style. These aspects make social services the most powerful sensor for events detection and automatic news generation. The aim of this workshop was to ask researchers to enter into such view, by studying how social platforms can be used in real-time scenarios to detect emerging events and enrich them with contextual information.

First, we would like to thank the organizing committee of LREC 2016 for giving us the opportunity to organize the workshop. Second, we would like to thank our program committee members. And of course, we would like to thank all the authors of the workshop for submitting their research works and for their participation.

We hope you will enjoy the second edition of the Sideways workshop and the LREC Conference, and have a great time in Portoroz.

The call for papers attracted submissions from India, Europe, Africa, and the United States.

The program committee reviewed and accepted the following:

Venue or Track	*Reviewed*	*Accepted*	
Full Papers	5	3	60%
Short Papers	2	0	0%
Total	7	3	42,8%

We also encourage attendees to attend the invited talk presentation.

- *Challenges for using social media for early detection of T2DM* - Dane Bell - University of Arizona
- *Exploiting Social Media to Address Fundamental Human Rights Issues* - Udo Krushwitz, Ayman Al Helbawy, Massimo Poesio - University of Essex

- *Tracking and analyzing the "second life" of TV content: a media and social-driven framework* - Luca Vignaroli, Claudio Schifanella, K. Selcuk Candan, R. Pensa, Maria Luisa Sapino – RAI Research Center, Arizona State University, University of Turin

Putting together *Sideways 2016* was a team effort. We thank the authors for providing the content of the program and we are grateful to the program committee who worked very hard in reviewing papers and providing feedback for authors. Finally, we thank the hosting organization.

We hope that you will find this program interesting and that the workshop will provide you with a valuable opportunity to share ideas with other researchers and practitioners from institutions around the world.

Co-Chairs
Luigi Di Caro
Mario Cataldi
Claudio Schifanella

Tuesday 24 May, 9:00 – 13:00
Chairs: Mario Cataldi, Luigi Di Caro, Claudio Schifanella

Challenges for using social media for early detection of T2DM

Dane Bell, Daniel Fried, Luwen Huangfu, Mihai Surdeanu, Stephen Kobourov

Twitter and other social media data are utilized for a wide variety of applications such as marketing and stock market prediction. Each application and appropriate domain of social media text presents its own challenges and benefits. We discuss methods for detecting obesity, a risk factor for Type II Diabetes Mellitus (T2DM), from the language of food on Twitter on community data, the peculiarities of this data, and the development of individual-level data for this task.

Catching Events in the Twitter Stream: A showcase of student projects

Tim Kreutz and Malvina Nissim

A group of bachelor students in information science at the University of Groningen applied off-the-shelf tools to the detection of events on Twitter, focusing on Dutch. Systems were built in four socially relevant areas: sports, emergencies, local life, and news. We show that (i) real time event detection is a feasible and suitable way for students to learn and employ data mining and analysis techniques, while building end-to-end potentially useful applications; and (ii) even just using off-the-shelf resources for such applications can yield very promising results.

Exploiting Social Media to Address Fundamental Human Rights Issues

Udo Krushwitz, Ayman Al Helbawy, Massimo Poesio

App2Check: a Machine Learning-based system for Sentiment Analysis of App Reviews in Italian Language

Emanuele Di Rosa, Alberto Durante

Sentiment Analysis has nowadays a crucial role in social media analysis and, more generally, in analysing user opinions about general topics or user reviews about product/services, enabling a huge number of applications. Many methods and software implementing different approaches exist and there is not a clear best approach for Sentiment classification/quantification. We believe that performance reached by machine learning approaches is a key advantage to apply to sentiment analysis in order to reach a performance which is very close to the one obtained by group of humans, who evaluate subjective sentences such as user reviews. In this paper, we present the App2Check system, developed mainly applying supervised learning techniques, and the results of our experimental evaluation, showing that App2Check outperforms state-of-the-art research tools on user reviews in Italian language related to the evaluation of apps published to app stores.

Distrusting Science on Communication Platforms: Socio-anthropological Aspects of the Science-Society Dialectic within a Phytosanitary Emergency

Christian Colella

The work aims to investigate the conspiracy-like and pseudo-scientific beliefs arose in Salento during a spread of a plant disease that affected olive tree crops known as "OQDS" and, more generally, tries to analyze through a socio-anthropological perspective the communication biases into the dialectical relationship between scientific research and general public and how social media platforms act like a conceptual container for pseudo-scientific belief and distrust sentiments toward science research.

Tracking and analyzing the "second life" of TV content: a media and social-driven framework

Luca Vignaroli, Claudio Schifanella, K. Selcuk Candan, Ruggero Pensa, Maria Luisa Sapino

People on the Web talk about television. TV users’ social activities implicitly connect the concepts referred to by videos, news, comments, and posts. The strength of such connections may change as the perception of users on the Web changes over time. With the goal of leveraging users’ social activities to better understand how TV programs are perceived by the TV public and how the users’ interests evolve in time, in this work, a framework that allows to manage, explore and analyze the heterogeneous and dynamic data coming from different information sources which play a role in what we call the “second life” of TV content will be exposed

EMOT:
Emotions, Metaphors, Ontology and Terminology during Disasters

Date: 28 May 2016

ABSTRACTS

Editors:

Khurshid Ahmad, Stephen Kelly, Xiubo Zhang

Abstracts of the LREC 2016 Workshop
“EMOT – Emotions, Metaphors, Ontology and Terminology during Disasters”

28 May 2016 – Portorož, Slovenia

Edited by Khurshid Ahmad, Stephen Kelly, Xiubo Zhang

This project has received funding from the European Union’s Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 607691.

SLANDAIL

Workshop Programme

09:30 – 09:40	Introduction by Workshop Chair
	Khurshid Ahmad, *Emotion, Metaphor, Ontology and Terminology*
09:40 – 10:30	Emotions
	Grazia Busa, Alice Cravotta, *Detecting Emotional Involvement in Professional News Reporters: an Analysis of Speech and Gestures*
	Maria Spyropoulou, *Emotive and Terminological Content in Disaster Related Messages*
10:30 – 11:00	Coffee break
11:00 – 11:25	Metaphors
	Carl Vogel, *Emotion, Quantification, Genericity, Metaphoricity*
11:25 – 11:50	Ontology
	Antje Schlaf, Sabine Gründer-Fahrer, Patrick Jähnichen *Topics in Social Media for Disaster Management – A German Case Study on the Flood 2013*
11:50 – 12:40	Terminology
	Maria Teresa Musacchio, Raffaella Panizzon, Xiubo Zhang, Virginia Zorzi *A Linguistically-driven Methodology for Detecting Impending and Unfolding Emergencies from Social Media Messages*
	Xiubo Zhang, Raffaella Panizzon, Maria Teresa Musacchio, Khurshid Ahmad, *Terminology Extraction for and from Communications in Multi-disciplinary Domains*
12:40 – 13:00	Closing Session

Workshop Organizers

Khurshid Ahmad	Trinity College Dublin, Republic of Ireland
Stephen Kelly	Trinity College Dublin, Republic of Ireland
Xiubo Zhang	Trinity College Dublin, Republic of Ireland

Workshop Programme Committee

Khurshid Ahmad	Trinity College Dublin, Republic of Ireland
Gerhard Heyer	Leipzig University, Germany
Bodil Nistrup Madsen	Copenhagen Business School, Denmark
Maria Teresa Musacchio	University of Padova, Italy
Henrik Selsoe Sorensen	Copenhagen Business School, Denmark
Carl Vogel	Trinity College Dublin, Republic of Ireland

Preface

An unexpected event induces an emotive reaction and metaphorical use of language; natural disasters, a class of unexpected events, induces the use of disaster-specific terminology and ontological descriptions in addition to emotive/metaphorical use of language. Disasters are characterised equally well by the fact that the onset, duration, and aftermath of this event, all to a greater or lesser degree, involve greater demand of information in a situation where information is generally scarce. One has to use all modalities of communications, including written and spoken language, visual communications, and non-verbal communications especially gestures.

The authors contributing to this workshop have been working on how to specify, design and prototype disaster management systems that use social media, including microblogging and social networks, as one of the inputs. Their linguistics coverage includes English, German and Italian. Their focus is on extracting information from continuous data streams including text, speech and images. There are two papers on emotional expressions used in disasters: Busa and Cravotta have analysed gestures and speech of reporters working in areas threatened by natural disasters like floods and suggest that hand gestures and the modulation of voice is correlate with the severity of an impending disaster. Spyropoulou looks at the terminological and affect content of text messages and speech excerpts and finds that speech comprises more information about the sentiment of the public at large than, say, their text messages. Topic modelling is one of the essential techniques that can be used to automatically categorise the contents of a text data stream of messages – this technique uses machine learning and information extraction, and has been successfully used by Schlaf, Gründer-Fahrer and Jähnichen to analyse German social media texts, especially on Facebook and Twitter: they find that the machine discovered categories that correspond quite ‘naturally’ to the categories of texts used in disaster management – warnings before disasters, re-location information during disasters, and requests after the disasters.

Vogel presents a theoretical discussion of the relationship between emotions and metaphors, and how affect-based language, laden with sentiment, is used. The selective use of terminology and ontology play a key role in a methodology form detecting impending and current emergencies in social media streams which has been developed by Musacchio, Panizzon and Zorzi. Finally, we have a paper that deals with the automatic extraction of terminology and ontology from text especially social media by Zhang et al: These authors have designed, implemented and evaluated an ontology/terminology extraction system (CiCui).

The authors appear to be aware of the problems relating to the factual and ethical provenance of data, especially on social media, involving as it does issues such as privacy, dignity, copyright ownership, rumours and many other legal and ethical considerations.

The presenters wish to acknowledge the support of the EU FP7 Programme focussing on the impact of social media in emergencies. The research leading to these results has received funding from the European community’s Seventh Framework Programme under grant agreement No. 607691 (SLANDAIL 2014:2017)

Emotion

Time *09:40 – 10:30*

Detecting Emotional Involvement in Professional News Reporters: an Analysis Of Speech and Gestures

G. Busa, A. Cravotta

Abstract

This study is aimed to investigate the extent to which reporters' voice and body behaviour may betray different degrees of emotional involvement when reporting on emergency situations. The hypothesis is that emotional involvement is associated with an increase in body movements and pitch and intensity variation. The object of investigation is a corpus of 21 10-second videos of Italian news reports on flooding taken from Italian nation-wide TV channels. The gestures and body movements of the reporters were first inspected visually. Then, measures of the reporters' pitch and intensity variations were calculated and related with the reporters' gestures. The effects of the variability in the reporters' voice and gestures were tested with an evaluation test. The results show that the reporters vary greatly in the extent to which they move their hands and body in their reportings. Two gestures seem to characterise reporters' communication of emergencies: beats and deictics. The reporters' use of gestures partially parallels the reporters' variations in pitch and intensity. The evaluation study shows that increased gesturing is associated with greater emotional involvement and less professionalism. The data was used to create an ontology of gestures for the communication of emergency.

Emotive and Terminological Content in Disaster Related Messages

M. Spyropoulou

Abstract

As the availability of information during a disaster is low, research has started to focus on other modes of communication that can complement text in information extraction and sentiment analysis applications. This paper attempts an initial estimation on what kind of results we should expect to get from disaster related text and speech data.

Metaphor

Time *11:00 – 11:25*

Emotion, Quantification, Genericity, Metaphoricity

C. Vogel

Topics in Social Media for Disaster Management – A German Case Study on the Flood 2013

A. Schlaf, S. Gründer-Fahrer, P. Jähnichen

Abstract

The paper presents the results of a German case study on social media use during the flood 2013 in Central Europe. During this event, thousands of volunteers organized themselves via social media without being motivated or guided by professional disaster management. The aim of our research was to show and analyze the real potential of social media for disaster management and to enable the public organizations to get into touch with the people and take advantage as well as control of the power of social media. In our investigation, we applied state-of-the-art technology from *Natural Language Processing*, mainly topic modeling, to test and demonstrate its usefulness for computer-based media analysis in modern disaster management. At the same time, the analysis was a comparative study of social media content in context of a disaster. We found that Twitter played its most prominent part in the exchange of current factual information on the state of the event, while Facebook prevalently was used for emotional support and organization of volunteers help. Accordingly, social media are powerful not only with respect to their volume, velocity and variety but also come with their own content, language and ways of structuring information.

Terminology

Time *11:50 – 12:40*

A Linguistically-driven Methodology for Detecting Impending and Unfolding Emergencies from Social Media Messages

M. Musacchio, R. Panizzon, X. Zhang, V. Zorzi

Abstract

Natural disasters have demonstrated the crucial role of social media before, during and after emergencies (Haddow & Haddow 2013). Within our EU project Slándáil, we aim to ethically improve the use of social media in enhancing the response of disaster-related agen-cies. To this end, we have collected corpora of social and formal media to study newsroom communication of emergency management organisations in English and Italian. Currently, emergency management agencies in English-speaking countries use social media in different measure and different degrees, whereas Italian National Protezione Civile only uses Twitter at the moment. Our method is developed with a view to identifying communicative strategies and detecting sentiment in order to distinguish warnings from actual disasters and major from minor disasters. Our linguistic analysis uses humans to classify alert/warning messages or emer-gency response and mitigation ones based on the terminology used and the sentiment expressed. Results of linguistic analysis are then used to train an application by tagging messages and detecting disaster- and/or emergency-related terminology and emotive language to simulate human rating and forward information to an emergency management system.

Terminology Extraction for and from Communications in Multi-disciplinary Domains

X. Zhang, R. Panizzon, M. Musacchio, K. Ahmad

Abstract

Terminology extraction generally refers to methods and systems for identifying term candidates in a uni-disciplinary and uni-lingual environment such as engineering, medical, physical and geological sciences, or administration, business and leisure. However, as human enterprises get more and more complex, it has become increasingly important for teams in one discipline to collaborate with others from not only a non-cognate discipline but also speaking a different language. Disaster mitigation and recovery, and conflict resolution are amongst the areas where there is a requirement to use standardised multilingual terminology for communication. This paper presents a feasibility study conducted to build terminology (and ontology) in the domain of disaster management and is part of the broader work conducted for the EU project Slándáil (FP7 607691). We have evaluated CiCui (for Chinese name 词萃, which translates to *words gathered*), a corpus-based text analytic system that combine frequency, collocation and linguistic analyses to extract candidates terminologies from corpora comprised of domain texts from diverse sources. CiCui was assessed against four terminology extraction systems and the initial results show that it has an above average precision in extracting terms.

Multimodal Corpora:
Computer Vision and Language Processing (MMC 2016)

24 May 2016

ABSTRACTS

Editors:

Jens Edlund, Dirk Heylen, Patrizia Paggio

Workshop Programme

09:00 – 09:15 – Welcome!

09:15 – 10:30 – Oral Session 1
Emiel van Miltenburg: *Stereotyping and Bias in the Flickr30k Dataset*
István Szekrényes, Laszlo Hunyadi and Tamas Varadi: *The Multimodal HuComTech Corpus: Principles of Annotation and Discovery of Hidden Patterns of Behaviour*
Michael Amory and Olesya Kisselev: *The Annotation of Gesture Designed for Classroom Interaction*

10:30 – 11:00 Coffee break

11:00 – 12:15 – Oral Session 2
Minghao Yang, Ronald Böck, Dawei Zhang, Tingli Gao, Linlin Chao, Hao Li and Jianhua Tao: *"Do You Like a Cup of Coffee?" - The CASIA Coffee House Corpus*
Paul Hongsuck Seo and Gary Geunbae Lee: *A Corpus for a Multimodal Dialog System for Presentation Controls*
Michael Tornow, Martin Krippl, Svea Bade, Angelina Thiers, Julia Krüger, Ingo Siegert, Sebastian Handrich, Lutz Schega and Andreas Wendemuth: *Integrated Health and Fitness (iGF)-Corpus - ten-Modal Highly Synchronized Subject-Dispositional and Emotional Human Machine Interactions*

12:15 – 13:45 – Lunch break

13:45 – 15:00 – Oral Session 3
Claire Bonial, Taylor Cassidy, Susan Hill, Judith Klavans, Matthew Marge, Douglas Summers-Stay, Garrett Warnell and Clare Voss: *A Robotic Exploration Corpus for Scene Summarization and Image-Based Question Answering*
Anna Matamala and Marta Villegas: *Building an Audio Description Multilingual Multimodal Corpus: The VIW Project*
Jindřich Libovický and Pavel Pecina: *A Dataset and Evaluation Metric for Coherent Text Recognition from Scene Images*

15:00 – 16:00 – Posters and demos
Ian Wood: *Thinspiration and Anorexic Tweets*
Emer Gilmartin, Ketong Su, Yuyun Huang, Christy Elias, Benjamin R. Cowan and Nick Campbell: *Collecting a human-machine, human-human, and human-woz comparative social talk corpus*
Kristin Hagen, Janne Bondi Johannessen, Anders Nøklestad and Joel Priestley: *Search and Annotation Tools for Heritage Language Spoken Corpora*

16:00 – 16:30 Coffee break

16:30 – 17:45 – Oral Session 4
Patrice Boucher, Pierrich Plusquellec, Pierre Dufour, Najim Dehak, Patrick Cardinal and Pierre Dumouchel: *PHYSIOSTRESS: A Multimodal Corpus of Data on Acute Stress and Physiological Activation*
Dimitra Anastasiou and Kirsten Bergmann: *A Gesture-Speech Corpus on a Tangible Interface*
Costanza Navarretta: *Filled pauses, Fillers and Familiarity in Spontaneous Conversations*

17:45 – 18:00 – Concluding remarks

Workshop Organizers

Jens Edlund	KTH Royal Institute of Technology
Dirk Heylen	University of Twente
Patrizia Paggio	University of Copenhagen/University of Malta

Preface

The creation of a multimodal corpus involves the recording, annotation and analysis of several communication modalities such as speech, hand gesture, facial expression, body posture, gaze, etc. An increasing number of research areas have transgressed or are in the process of transgressing from focused single modality research to full-fledged multimodality research, and multimodal corpora are becoming a core research asset and an opportunity for interdisciplinary exchange of ideas, concepts and data.

Given this, we are pleased to present the 11th Workshop on Multimodal Corpora, once again be in the form of an LREC workshop.

As always, we aimed for a wide cross-section of the field, with contributions ranging from collection efforts, coding, validation, and analysis methods to tools and applications of multimodal corpora.

Success stories of corpora that have provided insights into both applied and basic research are welcome, as are presentations of design discussions, methods and tools. This year, we wanted to pay special attention to the integration of computer vision and language processing techniques – a combination that is becoming increasingly important as the accessible video and speech data increases, and the suggested theme for this instalment of Multimodal Corpora was how processing techniques for vision and language can be combined to manage, search, and process digital content.

This workshop follows similar events held at LREC 00, 02, 04, 06, 08, 10, ICMI 11, LREC 2012, IVA 2013, and LREC 2014.

Welcome!

Tuesday 24 May, 9:00 – 9:15

Oral Session 1

Tuesday 24 May, 9:15 – 10:30
Chairperson: TBD

Stereotyping and Bias in the Flickr30k Dataset

Emiel van Miltenburg

An untested assumption behind the crowdsourced descriptions of the images in the Flickr30k dataset (Young et al., 2014) is that they "focus only on the information that can be obtained from the image alone" (Hodosh et al., 2013, p. 859). This paper presents some evidence against this assumption, and provides a list of biases and unwarranted inferences that can be found in the Flickr30k dataset. Finally, it considers methods to find examples of these, and discusses how we should deal with stereotype-driven descriptions in future applications.

The Multimodal HuComTech Corpus: Principles of Annotation and Discovery of Hidden Patterns of Behaviour

István Szekrényes, Laszlo Hunyadi and Tamas Varadi

The HuComTech Corpus represents a detailed and extensive annotation of verbal and nonverbal human behaviour as manifested in formal and informal dialogues of more than 50 hours of audiovisual recordings. The participants were 110 university students, and the language of the dialogues in both settings was Hungarian. The initial aim of building the corpus was to acquire a wide range of data characteristic of human-human interaction in order to make generalisations for their implementation in more advanced human-machine interaction systems. We were especially interested in the formal and pragmatic ways of how dialogues are managed according to specific contexts. The view is becoming increasingly shared that, in order to make a conversation successful, the formal verbal, syntactic and semantic aspects of a conversation need to go hand in hand with its nonverbal aspects (the suprasegmentals of speech as well as a wide variety of gestures). It is especially important in a human-machine interaction, where a proper emphasis on multimodality can significantly add to the robustness of such systems: the recognition of a speaker's gestures by the machine agent can contribute to the generation of its contextually proper responses and, consequently, to the sense of cooperativeness, a crucial component of a successful interaction. The need to cooperate goes beyond understanding the propositional content of the verbal component that is enhanced by the gestural one: the participants (either human or machine) need to interpret the partner's intentions, as well as his/her emotions as complements to their actions and manifestations of reactions to such intentions. Therefore the annotation of the corpus was extended to those formal and pragmatic markers of the given dialogues that were considered both characteristic of human-human interactions and implementable in a human-machine interaction system.

The Annotation of Gesture Designed for Classroom Interaction

Michael Amory and Olesya Kisselev

In the past decade, the field of Applied Linguistics has witnessed an increased interest in the study of multimodal aspects of language and language acquisition, and the number of multimodal corpora that are designed to investigate classroom interactions, second language acquisition and second language pedagogy is on the rise. The promise is that these digital repositories of video-recordings will be able to take advantage of Corpus Linguistics tools and procedures in order to maximize and diversify analytical capabilities. However, the transcription conventions (i.e., annotation schemas) for multimodal features (such as gestures, gaze, and body movement) that are simple, systematic and searchable are not readily available. The current project focuses on developing an annotation schema for the transcription of gestures, integrating the research traditions Conversation Analysis, gesture research, ASL and Corpus Linguistics. The goal of the project is to create a set of conventions that have analytical and descriptive power required for gesture research but are manageable for the transcriber and reader to engage with and that are systematic to allow for searchability. The study utilizes video-recorded data from the Corpus of English for Academic and Professional Purposes developed at the Pennsylvania State University.

Coffee Break

Tuesday 24 May, 10:30 – 11:00

Oral Session 2

Tuesday 24 May, 11:00 – 12:15
Chairperson: TBD

"Do You Like a Cup of Coffee?" - The CASIA Coffee House Corpus

Minghao Yang, Ronald Böck, Dawei Zhang, Tingli Gao, Linlin Chao, Hao Li and Jianhua Tao

Virtual agents are perfect options to represent a technical device in an interaction. However an appearance may influence in particular the user's behaviour. Therefore, we present a corpus containing naturalistic communication between a user and two virtual agents in a coffee house environment. The scenario is set up in a pleasant way dealing with various topics like discussion on coffee, drinks, weather, and gaming. Thus, behavioural studies are feasible. The combination of the agents is inspired by the "Utah scenario" but fitting two screens which allows for a non-static behaviour of the user. The multimodally recorded dataset provides audio-visual material in Mandarin from more than 50 participants. Further, Kinect recordings and transcripts are available as well. Besides the corpus description first insights on the material are given. Based on the Kinect's data movement and behaviour analyses of users in a Human-Computer Interaction are possible. Currently we identified two prototypical movement patterns in this context. Furthermore, we investigated in parallel the user's movement, head direction information, and the dialogue course. An interesting finding is that in such an interaction a full turn of the head and body happens relatively late even if the turn already fully shifted between agents.

A Corpus for a Multimodal Dialog System for Presentation Controls

Paul Hongsuck Seo and Gary Geunbae Lee

Research studies on multimodal data recently received great interest. Especially combining information in human verbal and gestural modalities is rapidly emerging based on findings of literatures that say the verbal and gestural modalities are highly correlated to each other even in their production. However, there are not many available resources comprising the verbal and gestural modalities making a barrier to start research studies on practical applications. In this ongoing work, we aim to build a multimodal corpus comprising the verbal and gestural modalities designed for a dialog system for presentation controls having eight different features with 18 user intents. The collected data is very rich in modality containing active IR and HD colour videos with audio stream from four microphones array. We present the annotation process of utterances and gestures of the collected presentation recordings for such application with its statistics.

Integrated Health and Fitness (iGF)-Corpus - ten-Modal Highly Synchronized Subject-Dispositional and Emotional Human Machine Interactions

Michael Tornow, Martin Krippl, Svea Bade, Angelina Thiers, Julia Krüger, Ingo Siegert, Sebastian Handrich, Lutz Schega and Andreas Wendemuth

A multimodal corpus on human machine interaction in the area of health and fitness is introduced in this paper. It shows the interaction of users with a gait training system. The subjects pace through a training course four times. In the intermissions, they interact with a multimodal platform, where they are given feedback, they re-assess their performance and they plan the next steps. A high involvement of the subjects is given. By design, the interaction further evokes cognitive underload and overload and emotional reactions. The platform interaction was arranged as a Wizard of Oz Setup. In the interaction phase, 10 modalities are recorded in 20 sensory channels with high performance of hardware synchronicity, including several high-resolution cameras, headset and directional microphones, biophysiology, 3D data as well as skeleton and face detection information. In the corpus, 65 subjects are recorded in the interaction sessions for a total of 100 minutes per subject, including self-ratings from eight time points during the experiment. Addi- tionally, several questionnaires are available from all subjects, regarding personality traits, including technical and stress coping behavior.

Lunch

Tuesday 24 May, 12:15 – 13:45

Oral Session 3

Tuesday 24 May, 13:45 – 15:00
Chairperson: TBD

A Robotic Exploration Corpus for Scene Summarization and Image-Based Question Answering

Claire Bonial, Taylor Cassidy, Susan Hill, Judith Klavans, Matthew Marge, Douglas Summers-Stay, Garrett Warnell and Clare Voss

The focus of this research is the development of training corpora that will facilitate multimodal human-robot communication. The corpora consist of video and images recorded during a robot's exploration of an office building and several types of associated natural language text, gathered in four distinct annotation tasks. Each annotation task supports the development of training data for a foundational technology integral to facilitating multimodal communication with robots, including object recognition, scene summarization, image retrieval, and image-based question answering. Although each of these technology areas has been addressed in work on computer vision, our research examines the unique requirements of these technologies when the visual data is collected by a robot and analyzed from a first-person perspective. For our purposes, the robot must be able to provide efficient natural language summaries of what it is seeing and respond to natural language queries with visual data. Here, we describe the progress of this ongoing research thus far, including the robotic exploration data and the development of annotation tasks.

Building an Audio Description Multilingual Multimodal Corpus: The VIW Project

Anna Matamala and Marta Villegas

This paper presents an audio description multilingual and multimodal corpus developed within the VIW (Visual Into Words) Project. A short fiction film was created in English for the project and was dubbed into Spanish and Catalan. Then, 10 audio descriptions in Catalan, 10 in English and 10 in Spanish were commissioned to professional describers. All these data were annotated at two levels (cinematic and linguistic) and were analysed using ELAN. The corpus is an innovative tool in the field of audiovisual translation research which allows for comparative analyses both intralingually and interlingually. Examples of possible analyses are put forward in the paper.

A Dataset and Evaluation Metric for Coherent Text Recognition from Scene Images

Jindřich Libovický and Pavel Pecina

In this paper, we deal with extraction of textual information from scene images. So far, the task of Scene Text Recognition (STR) has only been focusing on recognition of isolated words and, for simplicity, it omits words which are too short. Such an approach is not suitable for further processing of the extracted text. We define a new task which aims at extracting coherent blocks of text from scene images with regards to their future use in natural language processing tasks, mainly machine translation. For this task, we enriched the annotation of existing STR benchmarks in English and Czech and propose a string-based evaluation measure that highly correlates with human judgement.

Posters and Demos

Tuesday 24 May, 15:00 – 16:00
Chairperson: TBD

Thinspiration and Anorexic Tweets

Ian Wood

The online presence of anorexics and other people with eating disorders, the “pro-ana” (pro-anorexia) movement, has received much attention both in the media and in eating disorder research, with much contention about harmful and beneficial effects on the individuals who take part in the community. Several micro-blogging platforms are used by this community, notably including Tumblr, Instagram and Twitter. I present a collection of Tweets containing a selection of hash tags used by people with eating disorders that were collected over a 3 year period from November 2012. Images are a very prominent and important part of the collected data, with 71% of the tweets containing images and many of those tweets containing few or no words. In this demonstration, I will present some overall statistics and initial analyses of this data set as well as opportunities for enriching the data through detecting image features specific to this context, providing examples of common image types and features and an explanation of their relevance as symbols of the "thin ideal" and as indicators of the psychology of tweet authors and how that relates to eating disorder research. A multimodal analysis of this data, with image features combined with text, promises to yield greater and more informative insights into the Twitter eating disorder and thinspiration community than text analysis alone.

Collecting a human-machine, human-human, and human-woz comparative social talk corpus

Emer Gilmartin, Ketong Su, Yuyun Huang, Christy Elias, Benjamin R. Cowan and Nick Campbell

This paper describes the motivation for and collection of a multimodal corpus of short social interactions in three conditions - human-human, human-machine, and human-WOZ. The corpus has been designed for use in multimodal data-driven analysis of engagement in HMI, modelling of dyadic social talk for use in social spoken dialogue applications, and comparison of human-machine and human-human dialogue. The corpus design closely follows that of a similar French language corpus and will be used to compare social talk in the two languages.

Search and Annotation Tools for Heritage Language Spoken Corpora

Kristin Hagen, Janne Bondi Johannessen, Anders Nøklestad and Joel Priestley

Spoken corpora come in many shapes and sizes, often with their own, unique requirements. This demonstration looks at work with heritage languages, using the Corpus of American Norwegian Speech to show some of the challenges and design considerations. Following a brief account of data collection, focus will be on automatic annotation and providing multimodal access to the corpus.

Coffee Break

Tuesday 24 May, 16:00 – 16:30

Oral Session 4

Tuesday 24 May, 16:30 – 17:45
Chairperson: TBD

PHYSIOSTRESS: A Multimodal Corpus of Data on Acute Stress and Physiological Activation

Patrice Boucher, Pierrich Plusquellec, Pierre Dufour, Najim Dehak, Patrick Cardinal and Pierre Dumouchel

This paper presents PHYSIOSTRESS, our data corpus of acute stress and physiological activation. It includes 26 experiments of acute stress based on the Trier Social Stress Test, and 24 experiments which combine a social task and a physical activity. These experiments were accomplished by 13 men and 13 women between 20 and 49 years old without identified cardiac disease, respiratory disease neither mental disease. The psychosocial situation of each participant was evaluated from 5 questionnaires including the Rosenberg Self-Esteem Scale, the Perceived Stress Scale, the Trier Inventory for the assessment of Chronic Stress, the International Physical Activity Questionnaire} and a socio-demographic questionnaire. During each experiment, we record a derivation of ECG, respiratory data, the momentum of the subject (from accelerometers), the internal sounds of the body, audio and videos of the subjects and the 3D movements of the subjects (from a Kinect Microsoft sensor). The level of stress of each subject (no stress, low stress, medium stress or high stress) is annotated according to three references including: the stress felt by the subject, the stress apparent (annotated by two observers) and the subject level of salivary cortisol. PHYSIOSTRESS is a rare corpus of acute stress which combines measurements of heart and respiration with annotations of the salivary cortisol level, namely a standard in medical research in the evaluation of acute stress.

A Gesture-Speech Corpus on a Tangible Interface

Dimitra Anastasiou and Kirsten Bergmann

This paper presents a corpus of hand gestures and speech which is created using a tangible user interface (TUI) for collaborative problem solving tasks. We present our initial work within the European Marie Curie project GETUI (GEstures in Tangible User Interfaces). This project involves mainly creating a taxonomy of gestures used in relation to a tangible tabletop which is placed at the Luxembourg Institute of Science and Technology (LIST). A preliminary user study showed that gesturing encourages the use of rapid epistemic actions by lowering cognitive load. Ongoing corpus collection studies provide insights about the impact of gestures on learning, collaboration and cognition, while also identify cultural differences of gestures.

Filled pauses, Fillers and Familiarity in Spontaneous Conversations

Costanza Navarretta

This paper presents a pilot study of the use of fillers, filled pauses and co-occurring gestures in Danish spontaneous conversations between two or three people who know each other well and in dyadic first encounters. Fillers, such as the English uh and um are very common in spoken language and previous research has indicated that they function as interaction management and/or discourse planning signals. In particular, filled pauses have been found to be very frequent when speakers have to express something challanging. Previous research has also found a correlation between the degree of familiarity of the conversants and the frequency of speech overlaps and feedback unimodal and multimodal signals. In this study, we investigate whether the frequency of fillers and filled pauses and the familiarity degree of the conversants are connected and we hypothesize that fillers and filled pauses will occur more frequently in the first encounters than in conversations between persons who know each other because the communicative setting is more challenging in the former case. The results of our study confirm this hupothesis. our study also shows that fillers and filled pauses co-occur with gestures more frequently in the first encounters than in the conversations between acquainted participants. This might be due to the fact that people who are familiar do not need to signal that they want to take or give the turn as strongly as people who do not know each other. However, since many factors can influence communication, these findings should be confirmed on more data.

Concluding Remarks

Tuesday 24 May, 17:45 – 18:00

7th Workshop on the Representation and Processing of Sign Languages:

Corpus Mining

28 May 2016

ABSTRACTS

Editors:

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

Workshop Programme

09:00 – 10:30	On-stage Session A: *Corpus Mining*
10:30 – 11:00	Coffee break
11:00 – 13:00	Poster Session B: *Corpora and Mining*
13:00 – 14:00	Lunch break
14:00 – 16:00	Poster Session C: *New Challenges for SL Corpora and Resources*
16:00 – 16:30	Coffee break
16:30 – 18:00	On-stage Session D: *SL Resources: Collaboration and Sharing*

Workshop Organizers

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

Workshop Programme Committee

Penny Boyes Braem	Center for Sign Language Research, Basel CH
Annelies Braffort	LIMSI/CNRS, Orsay FR
Onno Crasborn	Radboud University, Nijmegen NL
Athanasia-Lida Dimou	Institute for Language and Speech Processing, Athens GR
Sarah Ebling	Institute of Computational Linguistics, University of Zurich, Zurich CH
Eleni Efthimiou	Institute for Language and Speech Processing, Athens GR
Michael Filhol	CNRS–LIMSI, Université Paris-Saclay, Orsay FR
Stavroula-Evita Fotinea	Institute for Language and Speech Processing, Athens GR
Thomas Hanke	Institute of German Sign Language, University of Hamburg, Hamburg DE
Julie Hochgesang	Gallaudet University, Washington US
Matt Huenerfauth	Rochester Institute of Technology, Rochester, NY, USA
Hernisa Kacorri	City University New York (CUNY), New York, USA
Athanasios Katsamanis	Computer Vision, Speech Communication and Signal Processing Group, National Technical University of Athens, Athens GR
Jette Kristoffersen	Centre for Sign Language, University College Capital, Copenhagen DK
John McDonald	DePaul University, Chicago US
Johanna Mesch	Stockholm University, Stockholm SE
Carol Neidle	Boston University, Boston US
Rosalee Wolfe	DePaul University, Chicago US

Session A: Corpus Mining

Saturday 28 May, 09:00 – 10:30

Chairperson: Julie Hochgesang On-stage Session

The Importance of 3D Motion Trajectories for Computer-based Sign Recognition

Mark Dilsizian, Zhiqiang Tang, Dimitri Metaxas, Matt Huenerfauth and Carol Neidle

Computer-based sign language recognition from video is a challenging problem because of the spatiotemporal complexities inherent in sign production and the variations within and across signers. However, linguistic information can help constrain sign recognition to make it a more feasible classification problem. We have previously explored recognition of linguistically significant 3D hand configurations, as start and end handshapes represent one major component of signs; others include hand orientation, place of articulation in space, and movement. Thus, although recognition of handshapes (on one or both hands) at the start and end of a sign is essential for sign identification, it is not sufficient. Analysis of hand and arm movement trajectories can provide additional information critical for sign identification. In order to test the discriminative potential of the hand motion analysis, we performed sign recognition based exclusively on hand trajectories while holding the handshape constant. To facilitate this evaluation, we captured a collection of videos involving signs with a constant handshape produced by multiple subjects; and we automatically annotated the 3D motion trajectories. 3D hand locations are normalized in accordance with invariant properties of ASL movements. We trained time-series learning-based models for different signs of constant handshape in our dataset using the normalized 3D motion trajectories. Results show significant computer-based sign recognition accuracy across subjects and across a diverse set of signs. Our framework demonstrates the discriminative power and importance of 3D hand motion trajectories for sign recognition, given known handshapes.

Towards a Visual Sign Language Corpus Linguistics

Thomas Hanke

Visualisations have a long tradition in linguistics, as in many fields dealing with complex structure. New forms of representations have been introduced to Visual Linguistics in the recent past, e.g. to help the researcher find the needle in a haystack, i.e. corpus. Here we present visualisation services available in iLex making a combined corpus and lexical database visually accessible. While many approaches suggested for textual languages transfer to sign language data as well, others explore sign-specific structure, such as multi-dimensional concordances not being restricted to sequentiality. Experimental combinations of animated visualisation and image processing might support the researcher to compensate for incomplete high-quality (=manual) annotation. In the long run, we see the potential that visualisation and data manipulation go hand in hand, allowing future user interfaces that are less text-heavy than today's sign language annotation environments.

Using Sign Language Corpora as Bilingual Corpora for Data Mining: Contrastive Linguistics and Computer-assisted Annotation

Laurence Meurant, Anthony Cleve and Onno Crasborn

More and more sign languages nowadays are now documented by large-scale digital corpora. But exploiting sign language (SL) corpus data remains subject to the time consuming and expensive manual task of annotating. In this paper, we present an ongoing research that aims at testing a new approach to better mine SL data. It relies on the methodology of corpus-based contrastive linguistics, exploiting SL corpora as bilingual corpora. We present and illustrate the main improvements we foresee in developing such an approach: downstream, for the benefit of the linguistic description and the bilingual (signed - spoken) competence of teachers, learners and the users; and upstream, in order to enable the automatisation of the annotation process of sign language data. We also describe the methodology we are using to develop a concordancer able to turn SL corpora into searchable translation corpora, and to derive from it a tool support to annotation.

Session B: Corpora and Mining

Saturday 28 May, 11:00 – 13:00
Chairperson: Johanna Mesch Poster Session

Visualizing Lects in a Sign Language Corpus: Mining Lexical Variation Data in Lects of Swedish Sign Language

Carl Börstell and Robert Östling

In this paper, we discuss the possibilities for mining lexical variation data across (potential) lects in Swedish Sign Language (SSL). The data come from the SSL Corpus (SSLC), a continuously expanding corpus of SSL, its latest release containing 43307 annotated sign tokens, distributed over 42 signers and 75 time-aligned video and annotation files. After extracting the raw data from the SSLC annotation files, we created a database for investigating lexical distribution/variation across three possible lects, by merging the raw data with an external metadata file, containing information about the age, gender, and regional background of each of the 42 signers in the corpus. We go on to present a first version of an easy-to-use graphical user interface (GUI) that can be used as a tool for investigating lexical variation across different lects, and demonstrate a few interesting finds. This tool makes it easier for researchers and non-researchers alike to have the corpus frequencies for individual signs visualized in an instant, and the tool can easily be updated with future expansions of the SSLC.

Linking Lexical and Corpus Data for Sign Languages: NGT Signbank and the Corpus NGT

Onno Crasborn, Richard Bank, Inge Zwitserlood, Els van der Kooij, Anique Schüller, Ellen Ormel, Ellen Nauta, Merel van Zuilen, Frouke van Winsum and Johan Ros

How can lexical resources for sign languages be integrated with corpus annotations? We answer this question by discussing an increasingly frequent scenario for sign language resources, where the lexical data are stored in an online lexical database that may also serve as a sign language dictionary, while the annotation data are offline files in the ELAN Annotation Format (EAF). There is by now broad consensus on the need for ID-glosses in corpus annotation, which in turn requires having at least a list of ID-glosses with a description of the phonological form and meaning of the signs. There is less of a consensus on standards for glossing, on practices of sign lemmatisation, and on the types of information that need to be stored in the lexical database. This paper contributes to the establishment of standards for sign language resources by discussing how two data resources for Sign Language of the Netherlands (NGT) are currently being integrated, using the ELAN annotation software for corpus annotation and an adaptation of the Auslan Signbank software as a lexical database. We discuss some of the present relations between two large NGT data sets, and outline some future developments that are foreseen.

From a Sign Lexical Database to an SL Golden Corpus – the POLYTROPON SL Resource

Eleni Efthimiou, Evita Fotinea, Athanasia - Lida Dimou, Theodore Goulas, Panagiotis Karioris, Kyriaki Vasilaki, Anna Vacalopoulou and Michalis Pissaris

The POLYTROPON lexicon resource is being created in an attempt i) to gather and recapture already available lexical resources of Greek Sign Language (GSL) in an up-to-date homogeneous manner, ii) to enrich these resources with new lemmas, and iii) to end up with a multipurpose-multiuse resource which can be equally exploited in end user oriented educational/communication services and in supporting various SL technologies. The database that hosts the newly acquired resource, incorporates various SL oriented fields of information, including information on compounding, GSL synonyms, classifier qualities, lemma related senses, semantic groupings etc, and also lemma coding for their manual and non-manual articulation activity. It also provides linking of GSL and Modern Greek equivalent(s) lemma pairs to serve bilingual use purposes. A by-product of considerable value is the parallel corpus which derived from the GSL examples of use accompanying each lemma entry in the dictionary and their translations into Modern Greek. The annotation of the corpus for the entailed signs and assignment of respective glosses in combination with data capturing by both HD and Kinect cameras in three repetitions, allowed for the creation of a golden parallel corpus available to the community of SL technologies for experimentation with various approaches to SL recognition, MT and information retrieval.

Annotated Video Corpus on FinSL with Kinect and Computer-vision Data

Tommi Jantunen, Outi Pippuri, Tuija Wainio, Anna Puupponen and Jorma Laaksonen

This paper presents an annotated video corpus of Finnish Sign Language (FinSL) to which has been appended Kinect and computer-vision data. The video material consists of signed retellings of the stories Snowman and Frog, where are you?, elicited from 12 native FinSL signers in a dialogue setting. The recordings were carried out with 6 cameras directed toward the signers from different angles, and 6 signers were also recorded with one Kinect motion and depth sensing input device. All the material has been annotated in ELAN for signs, translations, grammar and prosody. To further facilitate research into FinSL prosody, computer-vision data describing the head movements and the aperture changes of the eyes and mouth of all the signers has been added to the corpus. The total duration of the material is 45 minutes and that part of it that is permitted by research consents is available for research purposes via the LAT online service of the Language Bank of Finland. The paper briefly demonstrates the linguistic use of the corpus.

Methods for Recognizing Interesting Events within Sign Language Motion Capture Data

Pavel Jedlička, Zdeněk Krňoul and Miloš Železný

Rising popularity of motion capture in movie-production makes this technology more robust and more accessible. Utilization of this technology for sign language capturing and analysis is evident. The article deals with the usability of the motion capture in creating sign language corpora. A large amount of the data acquired by the motion capture has to be processed to provide usable data for wide range of research areas: e.g. sign language recognition, translation, synthesis, linguistics, etc. The aim of this article is to explore possible methods to detect interesting events in data using machine learning techniques. The result is a method for detection of the beginning and the end of the sign, hand location, finger and palm orientation, whether the sign is one or two handed, and symmetry in the two-handed signs.

Centroid-Based Exemplar Selection of ASL Non-Manual Expressions using Multidimensional Dynamic Time Warping and MPEG4 Features

Hernisa Kacorri, Ali Raza Syed, Matt Huenerfauth and Carol Neidle

We investigate a method for selecting recordings of human face and head movements from a sign language corpus to serve as a basis for generating animations of novel sentences of American Sign Language (ASL). Drawing from a collection of recordings that have been categorized into various types of non-manual expressions (NMEs), we define a method for selecting an exemplar recording of a given type using a centroid-based selection procedure, using multivariate dynamic time warping (DTW) as the distance function. Through intra- and inter-signer methods of evaluation, we demonstrate the efficacy of this technique, and we note useful potential for the DTW visualizations generated in this study for linguistic researchers collecting and analyzing sign language corpora.

The Usability of the Annotation

Jarkko Keränen, Henna Syrjälä, Juhana Salonen and Ritva Takkinen

Several corpus projects for sign languages have tried to establish conventions and standards for the annotation of signed data. When discussing corpora, it is necessary to develop a way of considering and evaluating holistically the features and problems of annotation. This paper aims to develop a conceptual framework for the evaluation of the usability of annotations. The purpose of the framework is not to give conventions for annotating but to offer tools for the evaluation of the usability of the annotation, in order to make annotations more usable and make it possible to justify and explain decisions about annotation conventions. Based on our experience of annotation in the corpus project of Finland's Sign Languages (CFINSL), we have developed six principles for the evaluation of annotation. In this article, using these six principles, we evaluate the usability of the annotations in CFINSL and other corpus projects. The principles have offered benefits in CFINSL: We are able to evaluate our annotations more systematically and holistically than ever before. Our work can be seen as an effort to bring a framework of usability to corpus work.

Transitivity in RSL: A Corpus-based Account

Vadim Kimmelman

A recent typological study of transitivity Haspelmath (2015) demonstrated that verbs can be ranked according to transitivity prominence, that is, according to how likely they are to be transitive cross-linguistically. This ranking can be argued to be cognitively rooted (based on the properties of the events and their participants) or frequency-related (based on the frequency of different types of events in the real world). Both types of explanation imply that the transitivity ranking should apply across modalities. To test it, we analysed transitivity of frequent verbs in the corpus of Russian Sign Language by calculating the proportion of overt direct and indirect objects and clausal complements. We found that transitivity as expressed by the proportion of overt direct objects is highly positively correlated with the transitive prominence determined cross-linguistically. We thus confirmed the modality-independent nature of transitivity ranking.

Automatic Alignment of HamNoSys Subunits for Continuous Sign Language Recognition

Oscar Koller, Hermann Ney and Richard Bowden

This work presents our recent advances in the field of automatic processing of sign language corpora targeting continuous sign language recognition. We demonstrate how generic annotations at the articulator level, such as HamNoSys, can be exploited to learn subunit classifiers. Specifically, we explore cross-language-subunits of the hand orientation modality, which are trained on isolated signs of publicly available lexicon data sets for Swiss German and Danish sign language and are applied to continuous sign language recognition of the challenging RWTH-PHOENIX-Weather corpus featuring German sign language. We observe a significant reduction in word error rate using this method.

Designing a Lexical Database for a Combined Use of Corpus Annotation and Dictionary Editing

Gabriele Langer, Thomas Troelsgård, Jette Kristoffersen, Reiner Konrad, Thomas Hanke and Susanne König

In a combined corpus-dictionary project, you would need one lexical database that could serve as a shared "backbone" for both corpus annotation and dictionary editing, but it is not that easy to define a database structure that applies satisfactorily to both these purposes. In this paper, we will exemplify the problem and present ideas on how to model structures in a lexical database that facilitate corpus annotation as well as dictionary editing. The paper is a joint work between the DGS Corpus Project and the DTS Dictionary Project. The two projects come from opposite sides of the spectrum (one adjusting a lexical database grown from dictionary making for corpus annotating, one building a lexical database in parallel with corpus annotation and editing a corpus-based dictionary), and we will consider requirements and feasible structures for a database that can serve both corpus and dictionary.

A New Tool to Facilitate Prosodic Analysis of Motion Capture Data and a Datadriven Technique for the Improvement of Avatar Motion

John McDonald, Rosalee Wolfe, Ronnie Wilbur, Robyn Moncrief, Evie Malaia, Sayuri Fujimoto, Souad Baowidan and Jessika Stec

Researchers have been investigating the potential rewards of utilizing motion capture for linguistic analysis, but have encountered challenges when processing it. A significant problem is the nature of the data: along with the signal produced by the signer, it also contains noise. The first part of this paper is an exposition on the origins of noise and its relationship to motion capture data of signed utterances. The second part presents a tool, based on established mathematical principles, for removing or isolating noise to facilitate prosodic analysis. This tool yields surprising insights into a data-driven strategy for a parsimonious model of life-like appearance in a sparse key-frame avatar.

The French Belgian Sign Language Corpus. A User-Friendly Corpus Searchable Online

Laurence Meurant, Aurélie Sinte and Eric Bernagou

This paper presents the first large-scale corpus of French Belgian Sign Language (LSFB) available via an open access website (www.corpus-lsfb.be). Visitors can search within the data and the metadata. Various tools allow the users to find sign language video clips by searching through the annotations and the lexical database, and to filter the data by signer, by region, by task or by keyword. The website includes a lexicon linked to an online LSFB dictionary.

Sign Classification in Sign Language Corpora with Deep Neural Networks

Lionel Pigou, Mieke Van Herreweghe and Joni Dambre

Automatic and unconstrained sign language recognition (SLR) in image sequences remains a challenging problem. The variety of signers, backgrounds, sign executions and signer positions makes the development of SLR systems very challenging. Current methods try to alleviate this complexity by extracting engineered features to detect hand shapes, hand trajectories and facial expressions as an intermediate step for SLR. Our goal is to approach SLR based on feature learning rather than feature engineering. We tackle SLR using the recent advances in the domain of deep learning with deep neural networks. The problem is approached by classifying isolated signs from the Corpus VGT (Flemish Sign Language Corpus) and the Corpus NGT (Dutch Sign Language Corpus). Furthermore, we investigate cross-domain feature learning to boost the performance to cope with the fewer Corpus VGT annotations.

A Digital Moroccan Sign Language STEM Thesaurus

Abdelhadi Soudi and Corinne Vinopol

This paper presents a gesture-based linguistic approach to assisting Moroccan Sign Language (MSL) users in understanding and appropriately using Science, Technology, Engineering and Mathematics (STEM) terminology by creating the first-ever digital MSL STEM Thesaurus. The thesaurus enables Deaf individuals to describe signs and obtain Standard Arabic word equivalents, concept graphics, and definitions in both MSL and Arabic. This is accomplished not only by providing words comparable to signs that they know, but also by providing other information (e.g., signed definitions) that helps differentiate Arabic word choices. The thesaurus is supported by a Concordancer for better illustration and disambiguation of STEM terms. The thesaurus will likely prove to be an invaluable tool that will enable children and adults who rely on MSL for communication, both deaf and otherwise communication impaired, to better understand and write knowledgeably and clearly on STEM topics, and pass standardized assessments.

Online Concordancer for the Slovene Sign Language Corpus SIGNOR

Špela Vintar and Boštjan Jerko

We present the first version of an online concordancing tool for the Slovene Sign Language SIGNOR corpus. The corpus search tool allows querying the SIGNOR annotated database by glosses and displays the hits in a keyword-in-context (KWIC) format, accompanied by frequency information, HamNoSys transcription and metadata. The main purpose of the tool is linguistic research, more specifically sign language lexicography, but also providing general public access to the corpus.

Session C: New Challenges for SL Corpora and Resources

Saturday 28 May, 14:00 – 16:00
Chairperson: Jette Kristoffersen Poster Session

The SIGNificant Chance Project and the Building of the First Hungarian Sign Language Corpus

Csilla Bartha, Margit Holecz and Szabolcs Varjasi

The Act CXXV of 2009 on Hungarian Sign Language and the Use of Hungarian Sign Language recognizes Hungarian Sign Language (HSL) as an independent natural language, moreover it provides the legal framework to introduce bilingual education (HSL-Hungarian) in 2017. In order to establish the linguistic background for bilingual education it was crucial to carry out linguistic research on HSL, which research should be sociolinguistically underpinned and should include corpus-based research. This research also aims to standardize HSL for educational purposes with the highest possible degree of community engagement. During the SIGNificant Chance project a sign language corpus (approximately 1750 hours) was created. A nation-wide fieldwork was conducted (five regions, nine venues). 147 sociolinguistic interviews and 27 grammatical tests (with 54 participants) were recorded in multiple-camera settings. There were also Hungarian competency tests and narrative interviews conducted with selected participants in order to make the complex description of their different linguistic practices in different discursive contexts possible. We are using ELAN and three different templates to analyze the collected data for different purposes (sociolinguistic-grammatical template, another for short term project purposes, and one for the dictionary). Some parts of the annotation work has been finished which contributed to the writing of the basic grammar of HSL and the creation of a small corpus-based dictionary of HSL.

Collecting and Analysing a Motion-Capture Corpus of French Sign Language

Mohamed-El-Fatah Benchiheub, Bastien Berret and Annelies Braffort

This paper presents a 3D corpus of motion capture data on French Sign Language (LSF), which is the first one available for the scientific community for pluridisciplinary studies. The paper also exhibits the usefulness of performing kinematic analysis on the corpus. The goal of the analysis is to acquire informative and quantitative knowledge for the purpose of better understanding and modelling LSF movements. Several LSF native signers are involved in the project. They were asked to describe 25 pictures in a spontaneous way while the 3D position of various body parts was recorded. Data processing includes identifying the markers, interpolating the information of missing frames, and importing the data to an annotation software to segment and classify the signs. Finally, we present the results of an analysis performed to characterize information-bearing parameters and use them in a data mining and modelling perspective.

Digging into Signs: Emerging Annotation Standards for Sign Language Corpora

Kearsy Cormier, Onno Crasborn and Richard Bank

This paper describes the creation of annotation standards for glossing sign language corpora as part of the Digging into Signs project (2014-2015). This project was based on the annotation of two major sign language corpora, the BSL Corpus (British Sign Language) and the Corpus NGT (Sign Language of the Netherlands). The focus of the gloss annotations in these data sets was in line with the starting point of most sign language corpora: to make general corpus annotation maximally useful regardless of the particular research focus. Therefore, the joint annotation guidelines that were the output of the project focus on basic annotation of hand activity, aiming to ensure that annotations can be made in a consistent way irrespective of the particular sign language. The annotation standard provides annotators with the means to create consistent annotations for various types of signs that in turn will facilitate cross-linguistic research. At the same time, the standard includes alternative strategies for some types of signs. In this paper we outline the key features of the joint annotation conventions arising from this project, describe the arguments around providing alternative strategies in a standard, as well as discuss reliability measures and improvement to annotation tools.

Recognition of Sign Language Hand Shape Primitives With Leap Motion

Burcak Demircioğlu, Güllü Bülbül and Hatice Köse

In this study, a rule based heuristic method is proposed to recognize the primitive hand shapes of Turkish Sign Language (TID) which are sensed by a Leap Motion device. The hand shape data set was also tested with selected machine learning method (Random Forest), and the results of two approaches were compared. The proposed system required less data than the machine learning method, and its success rate was higher.

Linking a Web Lexicon of DSGS Technical Signs to iLex

Sarah Ebling and Penny Boyes Braem

A website for a lexicon of Swiss German Sign Language equivalents of technical terms was developed several years ago using Flash technology. In the intervening years, the backend research database was migrated from FileMaker to iLex. Here, we report on the development of a web platform that provides access to the same technical signs by extracting the relevant information directly from iLex. This new platform has many advantages: New sets of signs for technical terms can be added or existing ones modified in iLex at any time, and changes are reflected in the web platform upon refreshing the browser. Just as importantly, the new platform can now also be accessed through all major mobile operating systems, as it does not rely on Flash. We describe how information on the glosses, keywords, videos of citation forms, status, and uses of the technical signs is represented in iLex and how the corresponding web platform was built.

Juxtaposition as a Form Feature - Syntax Captured and Explained rather than Assumed and Modelled

Michael Filhol and Mohamed Nassime Hadjadj

In this article, we report on a study conducted to further the design a formal grammar model (AZee), confronting it to the traditional notion of syntax along the way. The model was initiated to work as an unambiguous linguistic input for signing avatars, accounting for all simultaneous articulators while doing away with the generally assumed and separate levels of lexicon, syntax, etc. Specifically, the work presented here focused on juxtaposition in signed streams (a fundamental feature of syntax), which we propose to consider as a mere form feature, and use it as the starting point of data-driven searches for grammatical rules. The result is a tremendous progress in coverage of LSF grammar, and fairly strong evidence that our initial goal is attainable. We give concrete examples of rules, and a clear illustration of the recursive mechanics of the grammar producing LSF forms, and conclude with theoretical remarks on the AZee paradigm in terms of syntax, word/sign order and the like.

Examining Variation in the Absence of a 'Main' ASL Corpus: The Case of the Philadelphia Signs Project

Jami N. Fisher, Julie Hochgesang and Meredith Tamminga

The Philadelphia Signs Project emerged from the community's desire to document their local ASL variety, originating at the Pennsylvania School for the Deaf. This variety is anecdotally reported to be notably different from other ASL varieties. This project is founded upon the consistent observations of this marked difference. We aim to uncover what, if anything, makes the Philadelphia variety distinct from other varieties in the United States. Beyond some lexical items, it is unknown what linguistic features mark this variety as "different." Comparison to other ASL varieties is difficult given the absence of a main and representative ASL corpus. This paper describes our sociolinguistic data collection methods, annotation procedures, and archiving approach. We summarize several preliminary observations about potentially dialect-specific features beyond the lexicon, such as unusual phonological alternations and word orders. Finally, we outline our plans to test these features with surveys for non-Philadelphians using Philadelphia lexical items, extending to more abstract phonological and syntactic features. This line of inquiry supplements our current archiving practices, facilitating comparison with a main corpus in the future. We maintain that even without a main corpus for comparison, it is essential to document a language variety when the community wishes to preserve it.

Slicing your SL data into Basic Discourse Units (BDUs). Adapting the BDU model (syntax + prosody) to Signed Discourse

Sílvia Gabarró-López and Laurence Meurant

This paper aims to propose a model for the segmentation of signed discourse by adapting the Basic Discourse Units (BDU) Model. This model was conceived for spoken data and allows the segmentation of both monologues and dialogues. It consists of three steps: delimiting syntactic units on the basis of the Dependency Grammar (DG), delimiting prosodic units on the basis of a set of acoustic cues, and finding the convergence point between syntactic and prosodic units in order to establish BDUs. A corpus containing data from French Belgian Sign Language (LSFB) will be firstly segmented according to the principles of the DG. After establishing a set of visual cues equivalent to the acoustic ones, a prosodic segmentation will be carried out independently. Finally, the convergence points between syntactic and prosodic units will give rise to BDUs. The ultimate goal of adapting the BDU Model to the signed modality is not only to allow the study of the position of discourse markers (DMs) as in the original model, but also to give an answer to a controversial issue in SL research such as the segmentation of SL corpus data, for which a satisfactory solution has not been found so far.

Evaluating User Experience of the Online Dictionary of the Slovenian Sign Language

Ines Kozuh, Primož Kosec and Matjaž Debevc

The extensive use of mobile devices and tablets has resulted in an increasing need for the ubiquitous availability of different types of dictionaries online. The purpose of our study was to evaluate the user experience and usability of the online dictionary of the Slovenian sign language. Six Slovenian hearing non-signers were included in the study. While using the online dictionary, participants were asked to complete six tasks: searching for a letter, a word, written explanation of the word, thematic section and particular fairy tale, as well as completing the quiz. In addition, the participants evaluated the usability of the online dictionary with the System Usability Scale. The findings revealed that participants perceived the tasks "searching for the word" and "searching for the thematic section" to be the most difficult tasks and "completing the quiz" to be the easiest one. Regarding the time measured, the task "searching for the word" was the most time-consuming and the task "searching for the letter" was the least. This study provides insights into how Slovenian hearing users perceive using the online dictionary of the Slovenian sign language and could be the basis for future research with users of Slovenian sign language.

Semiautomatic Data Glove Calibration for Sign Language Corpora Building

Zdeněk Krňoul, Jakub Kanis, Miloš Železný and Luděk Müller

The article deals with a recording procedure for sign language dataset building mainly for avatar synthesis systems. Combined data glove and optical capture technique is considered. We present initial experiences with the motion capture data produced by the CyberGlove3 gloves and a set of new tools to ease the recording process, glove calibration and proper interpretation by the 3D model. It results in a more flexible solution for the sign language capture integrating manual glove calibration with an automatic initialization, time synchronization and high-resolution sensor readings.

"Non-tokens": When Tokens Should not Count as Evidence of Sign Use

Gabriele Langer, Thomas Hanke, Reiner Konrad and Susanne König

Lemmatised corpora consist of tokens as instantiations of signs (types). Tokens usually count as evidences of the signs' use. Frequency of tokens is an important criterion for the lexical status of a sign. In combination with metadata on the signers' sociolinguistic backgrounds such as age, gender, and origin these tokens can also be analysed for regional and sociolinguistic variation. However, corpora may also contain instances of sign use that do not reflect the sign use of the person uttering them. This is particularly true for metalinguistic discussions of signs, malformed signing and slips of the hand as well as other phenomena such as copying/repeating signs of the interlocutors or from stimulus material. In our presentation we list and discuss different kinds of sign use (tokens) that should either not be counted as proof of a sign type at all or at least not as evidence of regular sign use by that particular person. Examples of these "non-tokens" are either taken from the DGS Corpus or from uploaded video answers of the DGS Feedback. We also discuss some implications on how to annotate these cases.

Creating Corpora of Finland's Sign Languages

Juhana Salonen, Ritva Takkinen, Anna Puupponen, Henri Nieminen and Outi Pippuri

This paper discusses the process of creating corpora of the sign languages used in Finland, Finnish Sign Language (FinSL) and Finland-Swedish Sign Language (FinSSL). It describes the process of getting informants and data, editing and storing the data, the general principles of annotation, and the creation of a web-based lexical database, the FinSL Signbank, developed on the basis of the NGT Signbank, which is a branch of the Auslan Signbank. The corpus project of Finland´s Sign Languages (CFINSL) started in 2014 at the Sign Language Centre of the University of Jyväskylä. Its aim is to collect conversations and narrations from 80 FinSL users and 20 FinSSL users who are living in different parts of Finland. The participants are filmed in signing sessions led by a native signer in the Audio-visual Research Centre at the University of Jyväskylä. The edited material is stored in the IDA storage service produced by the CSC – IT Center for Science, and the metadata will be saved into CMDI metadata. Every informant is asked to sign a consent form where they state for what kinds of purposes their signing can be used. The corpus data are annotated using the ELAN tool. At the moment, annotations are created on the levels of glosses and translation.

Session D: SL Resources: Collaboration and Sharing

Saturday 28 May, 16:30 – 18:00
Chairperson: Thomas Hanke On-stage Session

Towards an Annotation of Syntactic Structure in the Swedish Sign Language Corpus

Carl Börstell, Mats Wiren, Johanna Mesch and Moa Gärdenfors

This paper describes on-going work on extending the annotation of the Swedish Sign Language Corpus (SSLC) with a level of syntactic structure. The basic annotation of SSLC in ELAN consists of six tiers: four for sign glosses (two tiers for each signer; one for each of a signer's hands), and two for written Swedish translations (one for each signer). In an additional step by Östling et al. (2015), all glosses of the corpus have been further annotated for parts of speech. Building on the previous steps, we are now developing annotation of clause structure for the corpus, based on meaning and form. We define a clause as a unit in which a predicate asserts something about one or more elements (the arguments). The predicate can be a (possibly serial) verbal or nominal. In addition to predicates and their arguments, criteria for delineating clauses include non-manual features such as body posture, head movement and eye gaze. The goal of this work is to arrive at two additional annotation tier types in the SSLC: one in which the sign language texts are segmented into clauses, and the other in which the individual signs are annotated for their argument types.

Preventing Too Many Cooks from Spoiling the Broth: Some Questions and Suggestions for Collaboration between Projects in iLex

Penny Boyes Braem and Sarah Ebling

Collaborative development of sign language resources is fortunately becoming increasingly common. In the spirit of collaboration, having one shared lexicon for sign language projects is a big advantage. However, this poses challenges to aspects pertaining to consistency of data, privacy of informants, and intellectual property. This contribution points out some problems that arise, especially if the common data comes from projects of different institutions. We describe what we have found to be a sustainable legal framework for our collaborative iLex corpus lexicon, giving an overview of the different kinds of partners involved in the creation and exploitation of a shared iLex corpus lexicon and providing our answers to the questions we faced along with an outlook for the future.

Community Input on Re-consenting for Data Sharing

Deborah Chen Pichler, Julie Hochgesang, Doreen Simons and Diane Lillo-Martin

Development of large sign language corpora is on the rise, and online sharing of such corpora promises unprecedented access to high quality sign language data, with significant time-saving benefits for sign language acquisition research. Yet data sharing also brings complex logistical challenges for which few standardized practices exist, particularly with regard to the protection of participant rights. Although some ethical guidelines have been established for large-scale archiving of spoken or transcribed language data, not all of these are feasible for sign language video data, especially given the relatively small and historically vulnerable communities from which sign language data are typically collected. Our primary focus is the process of re-consenting participants whose original informed consent did not address the possibility of sharing their video data. We describe efforts to develop ethically sound, community-supported practices for data sharing and archiving, summarizing feedback collected from two focus groups including a cross-section of community stakeholders. Finally, we discuss general themes that emerged from the focus groups, placing them in the wider context of similar discussions previously published by other researchers grappling with these same issues, with the goal of contributing to best-practices guidelines for data archiving and sharing in the sign language research community.

ISA-12:
12^{th} Joint ACL - ISO Workshop
on Interoperable Semantic Annotation

Saturday, May 28, 2016

ABSTRACTS

Editor:

Harry Bunt, Tilburg University

Workshop Programme

08.45 – 09:00 Registration
09:00 – 09:10 Opening

09:15 – 09:45 Claire Bonial, Susan Brown and Martha Palmer: *A Lexically-Informed Upper Level Event Ontology*
09:45 – 10:00 James Pustejovsky, Martha Palmer, Annie Zaenen, and Susan Brown: *Integrating VerbNet and GL Predicative Structures*
10:00 – 10:15 Elisabetta Jezek, Anna Feltracco, Lorenzo Gatti, Simone Magnolini and Bernardo Magnini: *Mapping Semantic Types onto WordNet Synsets*
10:15 – 10:30 Petya Osenova and Kiril Simov: *Cross-level Semantic Annotation of Bulgarian Treebank*

10:30 – 11:00 Coffee break

11:00 – 11:30 Ielka van der Sluis, Shadira Leito and Gisela Redeker: *Text-Picture Relations in Cooking Instructions*
11:30 – 12:00 Kiyong Lee: *An Abstract Syntax for ISOspace with its <moveLink> Reformulated*
12:00 – 12:15 James Pustejovsky, Kiyong Lee and Harry Bunt: *Proposed ISO Standard Amendment AMD 24617-7 ISOspace*

12:15 – 14:00 Lunch break, during which:
12:30 – 13:30 ISO TC 37/SC 4 Working groups 2 and 5 plenary meeting

14:00 – 14:30 Ludivine Crible: *Discourse Markers and Disfluencies: Integrating Functional and Formal Annotations*
14:30 – 15:00 Harry Bunt and Rashmi Prasad: *ISO DR-Core: Core Concepts for the Annotation of Discourse Relations*
15:00 – 15:15 Benjamin Weiss and Stefan Hillmann: *Feedback Matters: Applying Dialog Act Annotation to Study Social Attractiveness in Three-Party Conversations*
15:15 – 15:30 Andreea Macovei and Dan Cristea: *Time Frames: Rethinking the Way to Look at Texts*
15:30 – 15:45 Tuan Do, Nikhil Krishnaswamy, and James Pustejovsky: *ECAT: Event Capture Annotation Tool*

15:45 – 16:15 Tea break

16:15 – 16:45 Julia Lavid, Marta Carretero and Juan Rafael Zamorano: *Contrastive (English-Spanish) Annotation of Epistemicity in the MULTINOT Project: Preliminary Steps*
16:45 – 17:15 Elisa Ghia, Lennart Kloppenburg, Malvina Nissim and Paola Pietrandrea: *A Construction-Centered Approach to the Annotation of Modality*
17:15 – 17.30 Kyeongmin Rim: *MEA 2: Portable Annotation Tool for General Natural Language Use*

17.30 –17.31 ISA-12 Workshop Closing, followed by discussion of proposed new ISO activities:

17:31 – 17.45 James Pustejovsky and Kiyong Lee: *Proposal for New ISO activity PWI 24617-x ISOspaceSem*
17:45 – 18:00 James Pustejovsky: *Proposal for New ISO activity PWI 24617-x VoxML*

Workshop Organizers

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

Workshop Programme Committee

Jan Alexandersson	DFKI, Saarbrücken
Harry Bunt	Tilburg University
Nicoletta Calzolari	CNR-ILC, Pisa
Thierry Declerck	DFKI, Saarbrücken
Liesbeth Degand	UCL, Louvain-la-Neuve
David DeVault	USC, Playa Vista, CA
Alex Chengyu Fang	City University of Hong Kong
Robert Gaizauskas	University of Sheffield
Daniel Hardt	Copenhagen Business School
Koiti Hasida	Tokyo University
Elisabetta Jezek	University of Pavia
Michael Kipp	Augsburg University of Applied Sciences
Kiyong Lee	Korea University, Seoul
Philippe Muller	Université Paul Sabatier, Toulouse
Malvina Nissim	CLCG, University of Groningen
Volha Petukhova	Universität des Saarlandes, Saarbrücken
Paola Pietrandrea	Université de Tours and CNRS LLL
Andrei Popescu-Belis	IDIAP, Martigny
Laurent Prévot	Université Aix-Marseille
James Pustejovsky	Brandeis University, Waltham, MA
Laurent Romary	INRIA and Humboldt Universität Berlin
Ted Sanders	University of Utrecht
Manfred Stede	Universität Potsdam
Thorsten Trippel	University og Tübingen
Piek Vossen	Free University Amsterdam
Annie Zaenen	Stanford University
Sandrine Zufferey	Université de Fribourg

Session 1

9:00 – 10:30

A Lexically-Informed Upper Level Event Ontology

Claire Bonial, Susan Windisch Brown and Martha Palmer

This paper summarizes ISO 24617-8 (ISO DR-Core), a new part of the ISO SemAF framework for semantic annotation. Within this framework a range of standards is developed to support the interoperable annotation of semantic phenomena. The effort to develop a standard for the annotation of semantic relations in discourse is split into two parts, of which ISO 24617-8 concerns the first part, formulating desiderata for the annotation of discourse relations and providing clear definitions for a set of 'core' discourse relations, based on an analysis of a range of theoretical approaches and annotation efforts. Following the ISO principles for semantic annotation, an abstract syntax as well as a concrete XML-based syntax for annotations were defined, together with a formal semantics. Mappings are provided between the ISO core relations and various other annotation schemes.

Verb Meaning in Context: Integrating VerbNet and GL Predicative Structures

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

This paper reports on aspects of a new research project aimed at enriching VerbNet's predicative structures with representations and mechanisms from Generative Lexicon Theory. This involves the introduction of systematic predicative enrichment to the verb's predicate structure, including an explicit identification of the mode of opposition structure inherent in the predicate. In addition, we explore a GL-inspired semantic componential analysis over VerbNet classes, in order to identify coherent semantic cohorts within the classes.

Mapping Semantic Types onto WordNet Synsets

Elisabetta Jezek, Anna Feltracco, Lorenzo Gatti, Simone Magnolini and Bernardo Magnini

In this paper, we report the results of an experiment aimed at automatically mapping corpus-derived Semantic Types to WordNet synsets. The algorithm for the automatic alignment of Semantic Types with WordNet synsets relies on lexical correspondence, i.e. it performs an automatic alignment of Semantic Types labels with the corresponding WordNet entry nouns, when present (for example, the Semantic Type [[Activity]] is mapped to synsets containing the entry noun *activity#n*). In this way, 150 Types out of 180 are mapped automatically, while 30 gaps have to be resolved manually. Automatic mapping based on lexical correspondence, however, does not guarantee that the mapping is good, i.e. that the items which make up the extension of a certain Semantic Types match the set of hyponyms of the corresponding synset(s). An evaluation of 43 Semantic Types against a gold standard reveals that, for 30% of them, a manual revision is needed

.

Cross-level Semantic Annotation of Bulgarian Treebank

Petya Osenova and Kiril Simov

The paper focuses on the cross-level semantic annotation of BulTreeBank. It discusses the annotation of distinct lexemes as well as MultiWord Expressions with senses from the BTB Wordnet, valency frames dictionary, and DBPedia URIs and classes. Also, one important application of the semantically annotated treebank is discussed – namely, for improving the Knowledge-based Word Sense Disambiguation task via extraction of new semantic relations.

Session 2

11:00 – 12:15

Text-Picture Relations in Cooking Instructions

Ielka van der Sluis, Shadira Leito and Gisela Redeker

Like many other instructions, recipes on packages with ready-to-use ingredients for a dish combine a series of pictures with short text paragraphs. The information presentation in such multimodal instructions can be compact (either text or picture) and/or cohesive (text and picture). In an exploratory corpus study, 30 instructions for semi-prefabricated meals were annotated for text-picture relations. A slight majority of the 452 actions in the corpus were presented only textually. A third were presented in text and picture, indicating a moderate amount of cohesion. A minority of 31 actions (7%) were presented only pictorially, suggesting that the potential for compact multimodal presentation may be rather limited in these instructions.

An Abstract Syntax for ISOspace with its <moveLink> Reformulated

Kiyong Lee

ISOspace introduces the movement link, tagged `<moveLink>`, to annotate how motions are related to spatial entities in language. As pointed out in SemAF principles 9iso 24617-6), ISOspace overlaps SemAF-SR (ISO 24617-5), which treats semantic roles in general. It also fails to conform to the *link* structure $<\eta, E, \rho>$ as formulated in Bunt et al. (2016). To resolve these problems, we first construct the general abstract syntax $\mathscr{A}\mathscr{S}yn$ of annotation structures on which the abstract syntax $\mathscr{A}\mathscr{S}yn_{isoSpace}$ of ISOspace is instantiated. Following the two axioms on motion-events and event-paths, discussed by Pustejovsky and Yocum (2013), we then propose to restore the event-path, introduced earlier by Pustejovsky et al. (l2010), as a genuine basic entity in the abstract syntax, while implementing it as such into a concrete syntax. We finally reformulate the movement link as relating the mover of a motion-event to an event-path, as triggered by that motion-event. We also illustrate how the newly formulated movement link (`<moveLink>`) interacts with the other links in ISOspace.

Session 3

14:00 – 15:45

Discourse Markers and Disfluencies: Integrating Functional and Formal Annotations
Ludivine Crible

While discourse markers (DMs) and (dis)fluency have been extensively studied in the past as independent phenomena, combining DM-level and disfluency-level annotations however has never been carried out before at a fine-grained level of informativeness. It is argued that the integration of formal and functional annotations, while facing a number of methodological and theoretical challenges, is not only possible and innovative (addressing the lack of consensus in the field) but also highly relevant to the investigation of form-meaning patterns. This paper reports the methodological aspects of an annotation protocol which integrates formal identification of (dis)fluency markers and a multi-layered description of discourse markers featuring, among others, semantic-pragmatic variables such as their domain and function in context. The challenges and merits of this integration are illustrated by a comparison of clustering tendencies between different functions of DMs in DisFrEn, a French-English comparable dataset. Quantitative results allow us to generate tentative interpretations of the relative fluency of some DM functions based on co-occurrence patterns, in line with a cognitive-functional approach to spoken language.

ISO DR-Core: Core Concepts for the Annotation of Disccourse Relations
Harry Bunt and Rashmi Prasad

This paper summarizes ISO 24617-8 (ISO DR-Core), a new part of the ISO SemAF framework for semantic annotation. Within this framework a range of standards is developed to support the interoperable annotation of semantic phenomena. The effort to develop a standard for the annotation of semantic relations in discourse is split into two parts, of which ISO 24617-8 concerns the first part, formulating desiderata for the annotation of discourse relations and providing clear definitions for a set of 'core' discourse relations, based on an analysis of a range of theoretical approaches and annotation efforts. Following the ISO principles for semantic annotation, an abstract syntax as well as a concrete XML-based syntax for annotations were defined, together with a formal semantics. Mappings are provided between the ISO core relations and various other annotation schemes.

Feedback Matters: Applying Dialog Act Annotation to Study Social Attractiveness in Three-Party Conversations
Benjamin Weiss and Stefan Hillmann

The relationship between verbal behavior and social attractiveness ratings are studied based on three-party conversation scenarios. The recorded conversations are annotated according to ISO 24617-2:2012, applying 11 classes. Intra-group likability ratings given by each interlocutor are correlated with frequencies of each dialog-act class. A linear model shows significant relations between likability ratings given to interlocutors and frequencies of three dialog act classes uttered by the rater. Two classes "positive" and "negative allo-feedback" are negatively related to likability, whereas "positive auto-feedback" shows a positive relation. An effect for the receivers side was not found. All participants met briefly before starting the experiment and also conducted a training conversation, which is why no assumption on cause and effect have been made. This exploratory study motivates to look deeper into the interdependence between verbal behavior and social relationships than just on surface features as speaking time and number of turns.

Time Frames: Rethinking the Way to Look at Texts

Andreea Macovei and Dan Cristea

In this paper, we discuss the challenging task of identifying time frames that may intersect in a text (a novel, a news article, a Facebook post, etc.), in a form more or less visible for the reader. By time frame, we mean a sequence of events or statements that an author exposes voluntarily; these time frames can be considered specific writing techniques where diverse narrative threads are used for the purpose of capturing the reader's attention regarding the story as it develops. A particularity of time frames is the fact that the transition from one time frame to another one seems to be rather difficult to discern and put in evidence by a forewarned annotator, while the consequences of the temporal discontinuities are understood naturally by a casual reader of the text. We are going to explain this notion and to determine if it is necessary to propose a remodelled temporal annotation for this issue.

Session 4

Date / Time *16:15 – 17:30*

ECAT: Event Capture Annotation Tool

Tuan Do, Nikhil Krishnaswamy, and James Pustejovsky

This paper introduces the Event Capture Annotation Tool (ECAT), a user-friendly interface tool for annotating events and their participants in video, capable of extracting 3D positional and orientational data about objects in video captured by Microsoft's Kinect hardware. The modeling language VoxML (Pustejovsky and Krsihnaswamy, 2016) underlies ECAT's object, program, and attribute representations, although ECAT uses its own spec for explicit labeling of motion instances. The demonstration will show the toool's workflow and the options available for capturing event-participant relations and browsing visual data. Mapping ECAT's output to VoxML will also be addressed.

Contrastive (English- Spanish) Annotation of Epistemicity in the MULTINOT Project: Preliminary Steps

Julia Lavid, Marta Carretero and Juan Rafael Zamorano

In this paper we describe the preliminary steps undertaken for the annotation of the conceptual domain of epistemicity in English and Spanish, as part of a larger annotation effort of modal meanings in the context of the MULTINOT project. These steps focus on: a) the instantiation of existing linguistic theories in the area of epistemicity, identifying and defining the categories to be used as tags for annotation; b) the design of an annotation scheme which captures both the functional-semantic dimension of epistemicity, on the one hand, and the language-specific realisations of epistemic meanings in both languages, on the other. These two dimensions are shown to be necessary for investigating relevant contrasts between English and Spanish in the area of epistemicity and for the large-scale annotation of comparable and parallel texts belonging to different registers in English and Spanish.

A Construction-Centered Approach to the Annotation of Modality

Elisa Ghia, Lennart Kloppenburg, Malvina Nissim and Paola Pietrandrea

We propose a comprehensive annotation framework for modality, which encompasses and supports existing annotation schemes, by adopting a construction-centered view. Rather than seeing modality as a feature of a trigger or of a target, we view it as a feature of the triad “trigger-target-relation”, which we name *construction*. We motivate the need for such an approach from a theoretical perspective, and we also show that a construction-centered annotation scheme is operationally valid. We evaluate inter-annotator agreement via a pilot study, and find that modalised constructions identified by different annotators can be successfully aligned, as a first crucial step towards further agreement evaluations.

Kyeongmin Rim: MEA 2: Portable Annotation Tool for General Natural Language Use

A pair of general-purpose annotation/adjudication tools, MAE and MAI, has been available for years and has successfully proven itself useful in many semantic annotation projects. We are releasing a newer version, MAE2, that inherits the original pair’s strengths of being adaptable, flexible, and portable. The new version is enhanced with new features to help rapid prototyping of the design of natural language annotation tasks, naturally modeling complex semantic structures, setting up a more focusable and consistent annotation work-flow, and assuring the quality of annotations. Also, as an open source project, to make it easier to modify the software for specialized features for a specific annotation task, the MAE2 has adopted common software design patterns.

2nd Workshop on Natural Language Processing for Translation Memories (NLP4TM 2016)

28 May 2016

ABSTRACTS

Editors:

Constantin Orasan, Carla Parra, Eduard Barbu, Marcello Federico

Workshop Programme

9:00 – 10:30: Session 1: Invited talks

9:00 – 9:30
Marcello Federico, *Machine translation adaptation from translation memories in ModernMT*

9:30 – 10:00
Núria Bel, *Data fever in the 21st century. Where to mine Language Resources*

10:00 – 10:30
Samuel Läubli, *Data, not Systems: a Better Way to Conduct the Business of Translation*

10:30 – 11:00 Coffee break

11:00 – 12:00: Session 2: Research papers

11:00 – 11:20
A. Bellandi, G. Benotto, G. Di Segni, E. Giovannetti, *Investigating the Application and Evaluation of Distributional Semantics in the Translation of Humanistic Texts: a Case Study.*

11:20 – 11:40
Tapas Nayak, Santanu Pal, Sudip Kumar Naskar, Sivaji Bandyopadhyay, Josef van Genabith, *Beyond Translation Memories: Generating Translation Suggestions based on Parsing and POS Tagging*

11:40 – 12:00
Friedel Wolff, Laurette Pretorius, Loïc Dugast, Paul Buitelaar, *Methodological pitfalls in automated translation memory evaluation*

12:00 – 13:30: Session 3: Cleaning of translation memories shared task

12:00 – 12:20
Eduard Barbu, Carla Parra Escartín, Luisa Bentivogli, Matteo Negri, Marco Turchi, Marcello Federico, Luca Mastrostefano, Constantin Orasan, *1st Shared Task on Automatic Translation Memory Cleaning Preparation and Lessons Learned*

12:20 – 13:00
Presentations of the systems that took part in the shared task

13:00 – 13:30
Round table

Workshop Organizers

Constantin Orasan	University of Wolverhampton, UK
Carla Parra	Hermes, Spain
Eduard Barbu	Translated, Italy
Marcello Federico	FBK, Italy

Workshop Programme Committee

Juanjo Arevalillo	Hermes, Spain
Yves Champollion	WordFast, France
Gloria Corpas	University of Malaga, Spain
Maud Ehrmann	EPFL, Switzerland
Kevin Flanagan	Swansea University, UK
Corina Forascu	University “Al. I. Cuza”, Romania
Gabriela Gonzalez	eTrad, Argentina
Rohit Gupta	University of Wolverhampton, UK
Manuel Herranz	Pangeanic, Spain
Samuel Läubli	Autodesk, Switzerland
Liangyou Li	DCU, Ireland
Qun Liu	DCU, Ireland
Ruslan Mitkov	University of Wolverhampton, UK
Aleksandros Poulis	Lionbridge, Sweden
Gabor Proszeky	Morphologic, Hungary
Uwe Reinke	Flensburg University of Applied Sciences, Germany
Michel Simard	NRC, Canada
Mark Shuttleworth	UCL, UK
Masao Utiyama	NICT, Japan
Mihaela Vela	Saarland University, Germany
Andy Way	DCU, Ireland
Joern Wuebker	Lilt, United States
Marcos Zampieri	Saarland University and DFKI, Germany

Preface

Translation Memories (TM) are amongst the most used tools by professional translators, if not the most used. The underlying idea of TMs is that a translator should benefit as much as possible from previous translations by being able to retrieve how a similar sentence was translated before. Moreover, the usage of TMs aims at guaranteeing that new translations follow the client's specified style and terminology. Despite the fact that the core idea of these systems relies on comparing segments (typically of sentence length) from the document to be translated with segments from previous translations, most of the existing TM systems hardly use any language processing for this. Instead of addressing this issue, most of the work on translation memories focused on improving the user experience by allowing processing of a variety of document formats, intuitive user interfaces, etc.

The term second generation translation memories has been around for more than ten years and it promises translation memory software that integrates linguistic processing in order to improve the translation process. This linguistic processing can involve matching of subsentential chunks, edit distance operations between syntactic trees, incorporation of semantic and discourse information in the matching process. Terminologies, glossaries and ontologies are also very useful for translation memories, by facilitating the task of the translator and ensuring a consistent translation. The field of Natural Language Processing (NLP) has proposed numerous methods for terminology extraction and ontology extraction. The building of translation memories from corpora is another field where methods from NLP can contribute to improving the translation process.

We are happy we could include in the workshop programme four contributions dealing with the aforementioned issues. In addition, the programme of the workshop in complemented by the presentations of three well-known researchers.

The first edition of this workshop organised at RANLP 2015 confirmed the fact that there is interest in the research community for the topics proposed. In addition, it highlighted the need for automatic methods for cleaning translation memories. For this reason, the second edition of the NLP4TM workshop also organises a shared task on cleaning translation memories in an attempt to make the creation of resources for translation memories easier.

The Organising Committee would like to thank the Programme Committee, who responded with very fast but also substantial reviews for the workshop programme. This workshop would not have been possible without the support received from the EXPERT project (FP7/2007-2013 under REA grant agreement no. 317471, http://expert-itn.eu).

Session 1: Invited talks

Friday 27 May, 9:00 – 10:30

Machine translation adaptation from translation memories in ModernMT

Marcello Federico, FBK Trento, Italy

Adapting machine translation systems to specific customers or domains usually requires long time and extensive effort in training and optimizing the system. ModernMT sets out to do away with this by developing a new MT technology that seamlessly integrate translation memories into the MT system and train it on the fly without any disruption of service nor any user intervention. ModernMT is a new MT technology funded by the European Union that will overcome the typical limitations of traditional MT systems. From the software engineer perspective, ModernMT will provide an easy to install, fast to train, and simple to scale platform, capable to simultaneously serve tens of thousands of translators working simultaneously. Users will be able to integrate ModernMT in their favorite CAT tool, such as MateCat and Trados Studio, via a plugin that will merge translation memory and machine translation functions. In particular, by uploading and connecting their private translation memory to ModernMT, and by updating it during their work, they will not only receive better matches, but also more appropriate machine translation suggestions, that will significantly enhance their user experience and productivity. Real-time training and adaptation of machine translation from multiple translation memories, however, requires very efficient processing, e.g. for text cleaning, tokenisation, tag management, word alignment, adaptation, etc. In my talk I will present the overall ModernMT architecture, discuss its development roadmap and report preliminary results.

Data fever in the 21st century. Where to mine Language Resources.

Núria Bel, Universitat Pompeu Fabra, Spain

Language Resources, especially parallel corpora and bilingual glossaries, are raw materials for a number of application tasks and are considered a critical supply for Natural Language Processing-based applications such as Machine Translation. More and more efforts are being made with the aim of finding and exploiting deposits of multilingual data. The web is considered the most obvious mine: special crawlers are devised to find multilingual webs from where to extract parallel corpora. But other sources are also being explored such as open data. Open data is a quite promising source of data, particularly in the case of public administration data, although some issues concerning access and formats must be taken into consideration. Moreover, Linked Open Data has also proved to be useful for producing multilingual glossaries.

In this presentation, I will review different initiatives to mine Multilingual Language Resources which might be of interest for producing domain-specific Translation Memories for different language pairs. Besides the creation of new Translation Memories, these resources may also be used for enriching, curating or quality assuring already existing ones.

Data, not Systems: a Better Way to Conduct the Business of Translation

Samuel Läubli, Autodesk Development S.à.r.l., Switzerland

Over the past two decades, Autodesk has been acquiring large volumes of professional translations through localizing software products into more than 20 languages. Besides classical translation memory leveraging, we use this data for providing natural language processing services such as full text search, terminology harvesting, or domain-specific statistical machine translation.

While these services have been shown to positively impact translation quality and/or throughput (e.g., Plitt & Masselot, 2010), the fact that they are normally accessed via purpose-built systems creates inefficiencies for providers and consumers alike. From a corporate perspective, the effort needed for their maintenance and support would be better spent on improving the services as such. Translators, on the other hand, lose time in familiarizing themselves with client-specific systems as well as switching between them and their usual translation workbench. Recent research suggests that bundling these components into mixed-initiative interfaces (Horvitz, 1999) makes translation much more efficient and rewarding (Green et al., 2015).

In this talk, I will detail our transition from providing translators with the data rather than the software we think they need.

Session 2: Research papers

Friday 27 May, 11:00 – 12:00

Investigating the Application and Evaluation of Distributional Semantics in the Translation of Humanistic Texts: a Case Study

A. Bellandi, G. Benotto, G. Di Segni, E. Giovannetti

Digital Humanities are persisting ascending and the need for translating humanistic texts using Computer Assisted Translation (CAT) tools demands for a specific investigation both of the available technologies and of the evaluation techniques. Indeed, humanistic texts can present deep differences from texts that are usually translated with CAT tools, due to complex interpretative issues, the request of heavy rephrasing, and the addition of explicative parts in order to make the translation fully comprehensible to readers and, also, stylistically pleasant to read. In addition, these texts are often written in peculiar languages for which no linguistic analysis tool can be available. We faced this situation in the context of the project for the translation of the Babylonian Talmud from Ancient Hebrew and Aramaic into Italian. In this paper we describe a work in progress on the application of distributional semantics to the informing of the Translation Memory, and on the evaluation issues arising from its assessment.

Beyond Translation Memories: Generating Translation Suggestions based on Parsing and POS Tagging

Tapas Nayak, Santanu Pal, Sudip Kumar Naskar, Sivaji Bandyopadhyay,
Josef van Genabith

This paper explores how translations of unmatched parts of an input sentence can be discovered and inserted into Translation Memory (TM) suggestions generated by a Computer Aided Translation (CAT) tool using a parse tree and part of speech (POS) tags to form a new translation which is more suitable for post-editing. CATaLog (Nayek et al., 2015) is a CAT tool based on TM and a modified Translation Error Rate (TER) (Snover et al., 2006) metric. Unmatched parts of the sentence to be translated can often be found in some other TM suggestions or in sentences which are not part of TM suggestions. Therefore, we can find the translations of those unmatched parts within the TM database itself. If we can merge the translations of the unmatched parts into one single sentence in a meaningful way, then post-editing effort will be reduced. Inserting the translations for the unmatched parts into TM suggestions may lead to loss of fluency in the generated target sentence. To avoid that, we use parsing and POS tagging together with a back off POS n-gram model to generate new translation suggestions.

Methodological pitfalls in automated translation memory evaluation

Friedel Wolff, Laurette Pretorius, Loïc Dugast, Paul Buitelaar

A translation memory system attempts to retrieve useful suggestions from previous translations to assist a translator in a new translation task. While assisting the translator with a specific segment, some similarity metric is usually employed to select the best matches from previously translated segments to present to a translator. Automated methods for evaluating a translation memory system usually use reference translations and also use some similarity metric. Such evaluation methods might be expected to assist in choosing between competing systems. No single evaluation method has gained widespread use; additionally the similarity metric used in each of these methods are not standardised either. This paper investigates the choice of fuzzy threshold during evaluation, and the consequences of different choices of similarity metric in such an evaluation method. Important considerations for automated evaluation of translation memory systems are presented.

Session 3: Shared task on cleaning of translation memories

Friday 27 May, 12:00 – 13:30

1st Shared Task on Automatic Translation Memory Cleaning Preparation and Lessons Learned

Eduard Barbu, Carla Parra Escartín, Luisa Bentivogli, Matteo Negri, Marco Turchi, Marcello Federico, Luca Mastrostefano, Constantin Orasan

This paper summarizes the work done to prepare the first shared task on automatic translation memory cleaning. This shared task aims at finding automatic ways of cleaning TMs that, for some reason, have not been properly curated and include wrong translations. Participants in this task are required to take pairs of source and target segments from TMs and decide whether they are right translations. For this first task three language pairs have been prepared: English – Spanish, English – Italian, and English – German. In this paper, we report on how the shared task was prepared and explain the process of data selection and data annotation, the building of the training and test sets and the implemented baselines for automatic classifiers comparison.

Controlled Language Applications Workshop (CLAW)

28 May 2016

ABSTRACTS

Editors:

Key-Sun Choi, Sejin Nam

Workshop Programme

14:00 – 14:15	Key-Sun Choi, Hitoshi Isahara, Kiyong Lee and Christian Galinski: *Introduction about ISO and CNL*
14:15 – 14:30	Hitoshi Isahara and Tetsuzo Nakamura: *Report from Japan*
14:30 – 14:55	Adam Wyner, Francois Lévy and Adeline Nazarenko: *An Underspecified Approach to a Controlled Language for Legal Texts - a Position Paper -*
14:55 – 15:20	Christian Galinski and Blanca Stella Giraldo Pérez: *Rule-Based Technical Writing: A Meta-Standard on Controlled Language Extended towards Controlled Communication*
15:20 – 15:45	Sylviane Cardey: *A Controlled Language for Sense Mining and Machine Translation for Applications in Mission-Critical Domains*
15:45 – 16:10	Christina Lohr and Robert Herms: *A Corpus of German Clinical Reports for ICD and OPS-based Language Modeling*
16:10 – 16:30	Coffee break
16:30 – 16:55	Xiaofeng Wu, Liangyou Li, Jinhua Du and Andy Way: *ProphetMT: Controlled Language Authoring Aid System Description*
16:55 – 17:30	Rei Miyata, Anthony Hartley, Cécile Paris and Kyo Kageura: *Evaluating and Implementing a Controlled Language Checker*
17:30 – 17:40	Closing

Workshop Organizers

Key-Sun Choi	KAIST
Hitoshi Isahara	Toyohashi University of Technology
Christian Galinski	Infoterm
Andy Way	Dublin City University
Teruko Mitamura	Carnegie Mellon University

Workshop Programme Committee

Hitoshi Isahara	Toyohashi University of Technology
Andy Way	Dublin City University
Christian Galinski	Infoterm
Teruko Mitamura	Carnegie Mellon University
Kiyong Lee	ISO/TC37
Key-Sun Choi	KAIST

Preface

Following the highly successful workshops on the Controlled Natural Language Simplifying Language Use at LREC2014, we are pleased to announce the 6th CLAW workshop, embracing an open range both of applications to standardizations, in conjunction with the 10th edition of the Language Resources and Evaluation Conference (LREC2016), 23-28 May 2016, Grand Hotel Bernardin Conference Center, Portorož, Slovenia.

This workshop will focus more on the issues like standardization toward the Controlled Language Applications and their related supporting research and implementation issues in cooperation with the controlled language application, ISO/TC37 standardization, and semantic web communities.

The workshop on the Controlled Language Applications invite papers for the current progress and results toward the standardizations of controlled language. This workshop also would like to encourage submissions on any of (but not limited to) the following topics: human communication protocols, controlled text authoring, conformance checking systems, controlled language authoring aids, memory-based authoring, (re-)authoring combined with translation, issues in Controlled Language design, industrial experience and evolving requirements, models, processing algorithm, terminology aspects, R&D projects, use case, related topics on summarization, question and answering, machine translation, quality and usability evaluations of controlled language.

The workshop will give equal emphasis to the academic, corporate and industrial perspectives, while bringing together researchers, developers, users, and potential users of controlled language systems from around the world. The goal of this workshop will be to bridge the gap between the theory, practice and applications of controlled language and to identify the existing and possible future controlled language applications, and what should be kept in a standard for controlled language application.

An Underspecified Approach to a Controlled Language for Legal Texts - a Position Paper -

Adam Wyner, Francois Lévy, Adeline Nazarenko

The texts of legislation and regulation must be structured and augmented in order to allow for semantic web services (querying, linking, and inference). However, it is difficult to accurately parse and semantically represent such texts due to conventional practices of the legal community, the length and complexity of legal language, and the textual ground of the law. Controlled natural languages have been proposed as an approach to adjust to the difficulties, where the source text is rewritten in some standard form. However, such an approach has not suited legal language due to its requirements and complexities, so standardization has been difficult to achieve. To navigate between the requirements and complexities of legal language, standardization, and a fully controlled natural language, we take a position to propose and exemplify an approach to a high-level controlled language, which is adapted to the legal domain, correlates with the source text, and also facilitates analysis for semantic web applications. The approach can make use of some available NLP processing tools.

Rule-Based Technical Writing: A Meta-Standard on Controlled Language Extended towards Controlled Communication

Christian Galinski, Blanca Stella Giraldo Pérez

Standardization of rule-based technical writing (RBTW) emerged in English in certain industries. It started with Simplified Technical English (STE), or Simplified English which is the original name of a controlled language standard originally developed for aerospace industry maintenance manuals. Formerly called AECMA Simplified English, ASD (Aerospace and Defence Industries Association of Europe) renamed it to ASD Simplified Technical English. ASD-STE became so widely used by other industries and for a wide range of document types, that ‘simplified English’ is often used as a generic term for ‘controlled language’. Today the controlled language approach is applied in probably about a hundred languages, particularly in user instructions of all sorts. Increasingly such user instructions have to be rendered eAccessible, as the Convention on the Rights of Persons with Disabilities (CPRD) has been adopted into national legislation by numerous countries. As the needs of persons with disabilities (PwD) should be taken into account, whether in paper form or on websites, a systematic approach is commended for the development of such content on paper and as equivalent web content. For this purpose, a meta-standard with rules for the formulation of RBTW guides or standards would be useful.

A Controlled Language for Sense Mining and Machine Translation for Applications in Mission-Critical Domains

Sylviane Cardey

In this paper we present methodologies as well as the theoretical contributions involving the analysis and generation of texts for the application of controlled languages in multilingual mission-critical domains particularly safety-critical such as aeronautics, medicine and civil protection, where reliable results are obligatory. We show that the analysis involves the extraction of sense, that is sense-mining, and the generation that of controlled texts and their machine translation. This work has involved language modelling based on micro-systemic linguistic analysis, this itself being underpinned by a formal mathematical model, and which also inherently provides traceability, mandatory in safety-critical applications. A norms based approach is described involving extraction and application of norms in order to use them in the methodologies, both for analysis and generation. Applications, application domains and applicability are discussed.

A Corpus of German Clinical Reports for ICD and OPS-based Language Modeling

Christina Lohr, Robert Herms

In the field of health care and corresponding clinical institutions all occurring treatments need to be registered and documented in a comprehensive manner. For clinical documentation, complex reports (e.g., surgical interventions) are dictated by doctors and subsequently typed by secretaries. These reports are annotated with standardized codes for diagnosed diseases (ICD) and executed procedures (OPS). In this paper, we present a corpus of 450 German written clinical reports constructed for evaluation purposes, in particular for language modeling. We investigated the potential of the hierarchical structures of ICD and OPS codes in order to construct content-based language models for the clinical context. Experimental results show that OPS-based language modeling performed best using the highest level of the corresponding standard.

ProphetMT: Controlled Language Authoring Aid System Description

Xiaofeng Wu, Liangyou Li, Jinhua Du, Andy Way

This paper presents ProphetMT, a monolingual Controlled Language (CL) authoring tool which allows users to easily compose an in-domain sentence with the help of tree-based SMT-driven auto-suggestions. The interface also visualizes target-language sentences as they are built by the SMT system. When the user is finished composing, the final translation(s) are generated by a tree-based SMT system using the text and structural information provided by the user. With this domain-specific controlled language, ProphetMT will produce highly reliable translations. The contributions of this work are: 1) we develop a user-friendly auto-completion-based editor which guarantees that the vocabulary and grammar chosen by a user are compatible with a tree-based SMT model; 2) by applying a shift-reduce-like parsing feature, this editor allows users to write from left-to-right and generates the parsing results on the fly. Accordingly, with this in-domain composing restriction as well as the gold-standard parsing result, a highly reliable translation can be generated.

Evaluating and Implementing a Controlled Language Checker

Rei Miyata, Anthony Hartley, Cécile Paris, Kyo Kageura

This paper describes the evaluation of a detection component of a controlled language (CL) checker designed to assist non-professional writers in creating Japanese source texts that conform to a set of writing rules. We selected 23 Japanese CL rules shown to be effective for two Japanese to English machine translation systems as well as human readability, and implemented them with simple pattern matching using an existing morphological analyser. To benchmark the performance of the component, we created a comprehensive test set of noncompliant and compliant sentences from Japanese municipal websites, with all rule violations manually annotated. The results showed that 15 rules achieved high F-measure scores (more than 0.8) with nine obtaining the score of 1.0, while the precision scores of eight rules are low (less than 0.7). A detailed analysis of the results indicated ways to improve performance. Finally, based on the evaluation, we created an interface designed to alleviate the low-precision issue and implemented a prototype CL checker that is operational online.

4REAL Workshop:
Workshop on Research Results Reproducibility and Resources Citation in Science and Technology of Language

28 May 2016

ABSTRACTS

Editors:

António Branco, Nicoletta Calzolari and Khlaid Choukri

Workshop Programme

09:00 – 10:30 Reproducibility

Luís Gomes, Gertjan van Noord, António Branco, Steve Neale, *Seeking to Reproduce "Easy Domain Adaptation"*

Kevin Cohen, Jingbo Xia, Christophe Roeder and Lawrence Hunter, *Reproducibility in Natural Language Processing: A Case Study of two R Libraries for Mining PubMed/MEDLINE*

Filip Graliński, Rafał Jaworski, Łukasz Borchmann and Piotr Wierzchoń, *Gonito.net - Open Platform for Research Competition, Cooperation and Reproducibility*

10:30 – 11:00 Coffee break

11:00 – 12:00 Citation

Jozef Milšutka, Ondřej Košarko and Amir Kamran, *SHORTREF.ORG - Making URLs Easy-to-Cite*

Gil Francopoulo, Joseph Mariani and Patrick Paroubek, *Linking Language Resources and NLP Papers*

12:00 – 13:00 Round table

Reproducibility in Language Science and Technology: Ready for the Integrity Debate?
Chair: António Branco

Workshop Organizers

António Branco	University of Lisbon
Nicoletta Calzolari	ILC-CNR / ELRA
Khalid Choukri	ELDA

Workshop Programme Committee

António Branco	University of Lisbon
Iryna Gurevych	Universität Darmstadt
Isabel Trancoso	INESC-ID / IST, University of Lisbon
Joseph Mariani	CNRS/LIMSI
Justus Roux	NWVU
Khalid Choukri	ELDA
Maria Gavrilidou	ILSP
Marko Grobelnik	Jozef Stefan Institute
Marko Tadic	University of Zagreb
Mike Rosner	University of Malta
Nicoletta Calzolari	ILC-CNR/ELRA
Nick Campbell	Trinity College Dublin
Senja Pollak	Jozef Stefan Institute
Stelios Piperidis	ILSP
Steven Krauwer	University of Utrecht
Thierry Declerck	DFKI
Torsten Zesch	University of Duisburg-Essen
Yohei Murakami	Language Grid Japan

Preface/Introduction

The discussion on research integrity has grown in importance as the resources allocated to and societal impact of scientific activities have been expanding (e.g. Stodden, 2013, Aarts *et al.*, 2015), to the point that it has recently crossed the borders of the research world and made its appearance in important mass media and was brought to the attention of the general public (e.g. Nail and Gautam, 2011, Zimmer, 2012, Begley and Sharon 2012, The Economist, 2013).

The immediate motivation for this increased interest is to be found in a number of factors, including the realization that for some published results, their replication is not being obtained (e.g. Prinz *et al.*, 2011; Belgley and Ellis, 2012); that there may be problems with the commonly accepted reviewing procedures, where deliberately falsified submissions, with fabricated errors and fake authors, get accepted even in respectable journals (e.g. Bohannon, 2013); that the expectation of researchers vis a vis misconduct, as revealed in inquiries to scientists on questionable practices, scores higher than one might expect or would be ready to accept (e.g. Fanelli, 2009); among several others.

Doing justice to and building on the inherent ethos of scientific inquiry, this issue has been under thorough inquiry leading to a scrutiny on its possible immediate causes and underlying factors, and to initiatives to respond to its challenge, namely by the setting up of dedicated conferences (e.g. WCRI – World Conference on Research Integrity), dedicated journals (e.g. RIPR – Research Integrity and Peer review), support platforms (e.g. COS – Center for Open Science), revised and more stringent procedures (e.g. Nature, 2013), batch replication studies (e.g. Aarts *et al.*, 2015), investigations on misconduct (e.g. Hvistendahl, 2013), etc.

This workshop seeks to foster the discussion and the advancement on a topic that has been so far given insufficient attention in the research area of language processing tools and resources (Branco, 2013, Fokkens *et al.*, 2013) and that has been an important topic emerging in other scientific areas. That is the topic of the reproducibility of research results and the citation of resources, and its impact on research integrity.

We invited submissions of articles that presented pioneering cases, either with positive or negative results, of actual replication exercises of previous published results in our area. We were interested also in articles discussing the challenges, the risk factors, the procedures, etc. specific to our area or that should be adopted, or adapted from other neighboring areas, possibly including of course the new risks raised by the replication articles themselves and their own integrity, in view of the preservation of the reputation of colleagues and works whose results are reported has having been replicated, etc.

By the same token, this workshop was interested also on articles addressing methodologies for monitoring, maintaining or improving citation of language resources and tools and to assess the importance of data citation for research integrity and for the advancement of natural language science and technology.

The present volume gathers the papers that were selected for presentation and publication after having received the sufficiently positive evaluation by three reviewers from the workshop's program committee.

We hope this workshop, collocated with the LREC 2016 conference, will help to open and foster the discussion on research results reproducibility and resources citation in the domain of science and technology of language.

18 April 2016

António Branco, Nicoletta Calzolari and Khalid Choukri

References:

Aarts, et al., 2015, “Estimating the Reproducibility of Psychological Science”, *Science*.

The Economist, 2013, "Unreliable Research: Trouble at the Lab", *The Economist*, October 19, 2013, online.

Begley, 2012, "In Cancer Science, Many “Discoveries” don’t hold up", *Reuters*, March 28th, online.

Begley and Ellis, 2012, "Drug Development: Raise Standards for Preclinical Cancer Research", *Nature*.

Bohannon, 2013, "Who’s Afraid of Peer Review?", *Science*.

Branco, 2013, "Reliability and Meta-reliability of Language Resources: Ready to initiate the Integrity Debate?", In *Proceedings of The 12th Workshop on Treebanks and Linguistic Theories* (TLT12).

COS, Centre for Open Science.

Fanelli, 2009, "How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data", *PL OS ONE*.

Fokkens, van Erp, Postma, Pedersen, Vossen and Freire, 2013, "Offspring from Reproduction Problems: What Replication Failure Teaches US", In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (ACL2013).

Hvistendahl, 2013, “China’s Publication Bazaar”, *Science*.

Nail, 2011, "Scientists’ Elusive Goal: Reproducing Study Results", *The Wall Street Journal*.

Nature, 2013, “Announcement: Reducing our Irreproducibility”, *Nature*, Editorial.

Prinz, Sclange and Asadullah, 2011, "Believe it or not: How much can We Rely on Published Data on Potential Drug Targets?", *Nature Reviews Drug Discovery* 10, 712.

RIPR, Research Integrity and Peer Review.

Stodden, 2013, "Resolving Irreproducibility in Empirical and Computational Research", *IMS Bulletin Online*.

WCRI, World Conference on Research Integrity.

Zimmer, 2012, "A Sharp Rise in Retractions Prompts Calls for Reform", *The New York Times*.

Session 1: Reproducibility

Saturday 28 May, 9:00 - 10:30

Seeking to Reproduce "Easy Domain Adaptation"

Luís Gomes, Gertjan van Noord, António Branco and Steve Neale

The frustratingly easy domain adaptation technique proposed by Daume III (2007) is simple, easy to implement, and reported to be ´very successful in a range of NLP tasks (named entity recognition, part-of-speech tagging, and shallow parsing), giving us high hopes of successfully replicating and applying it to an English↔Portuguese hybrid machine translation system. While our hopes became ‘frustration’ in one translation direction – as the results obtained with the domain-adapted model do not improve upon the in-domain baseline model – our results are more encouraging in the opposite direction. This paper describes our replication of the technique and our application of it to machine translation, and offers a discussion on possible reasons for our mixed success in doing so.

Reproducibility in Natural Language Processing: A Case Study of two R Libraries for Mining PubMed/MEDLINE

Kevin Cohen, Jingbo Xia, Christophe Roeder and Lawrence Hunter

There is currently a crisis in science related to highly publicized failures to reproduce large numbers of published studies. This work proposes, by way of case studies, a methodology for moving the study of reproducibility in computational work to a full stage beyond that of earlier work.
Specifically, it presents a case study in attempting to reproduce the reports of two R libraries for doing text mining of the PubMed/MEDLINE repository of scientific publications. The main findings are that a rational paradigm for reproduction of natural language processing papers can be established; the advertised functionality was difficult, but not impossible, to reproduce; and reproducibility studies can produce additional insights into the functioning of the published system. Additionally, the work on reproducibility lead to the production of novel user-centered documentation that has been accessed 260 times since its publication---an average of once a day per library.

Gonito.net - Open Platform for Research Competition, Cooperation and Reproducibility

Filip Graliński, Rafał Jaworski, Łukasz Borchmann and Piotr Wierzchoń

This paper presents the idea of applying an open source, web-based platform - Gonito.net - for hosting challenges for researchers in the field of natural language processing. Researchers are encouraged to compete in well-defined tasks by developing tools and running them on provided test data. The researcher who submits the best results becomes the winner of the challenge. Apart from the competition, Gonito.net also enables collaboration among researchers by means of source code sharing mechanisms. Gonito.net itself is fully open source, i.e. its source is available for download and compilation, as well as a running instance of the system, available at gonito.net. The key design feature of Gonito.net is using Git for managing solutions of problems submitted by competitors. This allows for research transparency and reproducibility.

Session 2: Citation

Saturday 28 May, 11:00 - 12:00

SHORTREF.ORG - Making URLs Easy-to-Cite

Jozef Milšutka, Ondřej Košarko and Amir Kamran

In this paper, we present an easy-to-cite and persistent infrastructure (shortref.org) for research and data citation in the form of a URL shortener service. The reproducibility of results is very important for the reuse of research and directly depends on the availability of research data. The advancements in web technologies made the redistribution of data much easier; however, due to the dynamic nature of the internet, the content is constantly on the move from one destination to another. The URLs researchers use for citing their work do not directly account for changes and when the users try to access the cited URLs, the URLs do not need to be working anymore. In our proposed solution, the shortened URLs are not basic URLs but use persistent identifiers and provide a reliable mechanism to make the target always accessible which can directly improve the impact of research.

Linking Language Resources and NLP Papers

Gil Francopoulo, Joseph Mariani and Patrick Paroubek

The Language Resources and Evaluation Map (LRE Map) is an accessible database on resources based on records collected during the submission of various major Natural Language Processing (NLP) conferences, including the Language Resources and Evaluation Conferences (LREC). The NLP4NLP is a very large corpus of scientific papers in the field of Speech and Natural Language Processing covering a large number of conferences and journals. In this article, we make the link between those two elements in order to study the mention of the LRE Map resource names within the NLP4NLP corpus.

Session 3: Round table

Saturday 28 May, 12:00 - 13:00

Reproducibility in Language Science and Technology: Ready for the Integrity Debate?

Chair: António Branco

Novel Incentives for Collecting Data and Annotation from People: types, implementation, tasking requirements, workflow and results

28 May 2016

ABSTRACTS

Editors:

Christopher Cieri

Workshop Programme

14:00 – 14:25 – Introduction by Workshop Chair

Christopher Cieri, *Novel Incentives in Language Resource Development*

14:25 – 15:15 – Novel Incentives in Data Collection and Requisite Processing

Nick Campbell, *Herme & Beyond; the Collection of Natural Speech Data*

Kensuke Mitsuzawa, Maito Tauchi, Mathieu Domoulin, Masanori Nakashima, Tomoya Mizumoto*, FKC Corpus: a Japanese Corpus from New Opinion Survey Service*

15:15 – 16:05 – Novel Incentives and Workflows for Annotation

Kara Greenfield, Kelsey Chan, Joseph P. Campbell, *A Fun and Engaging Interface for Crowdsourcing Named Entities*

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz and Chris Madge, *Novel Incentives for Phrase Detectives*

16:05 – 16:30 – Afternoon Coffee Break

16:30 – 17:20 – Understanding and Exploiting Data from Alternative Sources

Na'im Tyson, Jonathan Roberts, Jeff Allen, Matt Lipson, *Evaluation of Anchor Texts for Automated Link Discovery in Semi-structured Web Documents*

Maxine Eskenazi, Sungjin Lee, Tiancheng Zhao, Ting Yao Hu, Alan W Black, *Unconventional Approaches to Gathering and Sharing Resources for Spoken Dialog Research*

17:20 – 18:00 – The Future of Incentives, Workforces, Workflows and Data Exploitation

Mark Liberman, *Oral Histories: Linguistic Documentation as Social Media*

Wrap-Up and Discussion

Workshop Organizers

Christopher Cieri	Linguistic Data Consortium, University of Pennsylvania
Chris Callison-Burch	University of Pennsylvania
Nick Campbell	Trinity College Dublin
Maxine Eskenazi	Carnegie Mellon University
Massimo Poesio	University of Essex, University of Trento

Workshop Programme Committee

Christopher Cieri	Linguistic Data Consortium, University of Pennsylvania
Chris Callison-Burch	University of Pennsylvania
Nick Campbell	Trinity College Dublin
Maxine Eskenazi	Carnegie Mellon University
Massimo Poesio	University of Essex, University of Trento
Stephanie Strassel	Linguistic Data Consortium, University of Pennsylvania
Jonathan Wright	Linguistic Data Consortium, University of Pennsylvania

Preface/Introduction

Despite more than two decades of effort from many research groups and large data centers, the supply of LRs falls far short of need even in the languages with the greatest number of speakers, controlling the largest shares of the world economy. For languages with less international recognition, resources are scarce, fragmentary or absent. Recent programs such as DARPA LORELEI recognize and attempt to address this gap but even they will provide only core resources for a few dozen languages, a small proportion of the >7000 currently in use worldwide.

In Language Resource (LR) development the commonest incentives for contributors are monetary. Whether motivated by convenience or ethical beliefs, that bias limits the Human Language Technology (HLT) community's ability to collect data and understand how different incentives impact collection. Because linguistic innovation is effectively limitless, relying upon a limited resource, monetary compensation, to generate the data needed to document the world's language is certain to fall short. Instead LR developers and users must develop and employ incentives that scale beyond the budget of a 3- or 5-year program.

A few HLT projects have employed alternate incentives. *Phrase Detectives* provides entertainment, challenge and access to interesting reading in exchange for anaphora annotation. *Herme* gave participants the unusual experience of interacting verbally with a tiny, cute robot while recording their interactions. *Let's Go* mediates access to Pittsburgh Port Authority Transit bus schedules and route information while recording the interactions to improve system performance in real world situations especially for 'extreme' users such as non-native and elderly speakers.

However, outside our field, collections employ variable incentives to much greater effect, creating massive data resources. *LibriVox* offer contributors the chance to create audio recordings of classic works of literature, develop their skills as reader and voice actors, work within a community of similarly minded volunteers and enable access to the blind, illiterate and others. *Zooniverse* includes linguistic exercises such as the transcription of originally hand written bird watching journals and artists' diaries or of the typewritten labels of insect collections. Social media has employed a wide range of incentives including:

- access to information and entertainment
- possibilities for self-expression, sharing and publicizing intellectual or creative work
- chances to vent frustrations or convey thoughts, sometimes anonymously
- forums for socializing; exercises which develop competence that may lead to new prospects
- competition, status, prestige, and recognition
- payment or discounts in real and virtual worlds
- access to services and infrastructure based on contributions
- novel experiences and improved interactions, for example in a customer service encounter
- opportunities to contribute to a greater cause or good

While lagging behind in the use of novel incentives, HLT researchers have productively used crowdsourcing to lower collection and annotation costs and developed techniques for customizing tasking to meet the capacity of the crowd and fusing highly variable results into data sets that advance technology development. Similar techniques apply to the use of alternate incentives in collecting data from a non-traditional workforce.

This half-day workshop will open the discussion on incentives in data collection describing novel approaches and comparing with traditional monetary incentives. Related topics including: descriptions of projects that use the alternate incentives listed above or others; modifications of the data collection and annotation tasking or workflow to accommodate a new workforce, including crowdsourcing; techniques for exploiting the results of alternate incentives and novel workflows.

Introduction by Workshop Chair

Saturday 28 May, 14:00 – 14:25
Session Chairperson: Liberman

Novel Incentives in Language Resource Development

Christopher Cieri
The gap between supply of and demand for Language Resources continues to impede progress in linguistic research and technology development, even in the face of immense international effort to create the requisite data and tools. This deficiency affects all languages in some way, even those with worldwide economic and political influence, though for most of the world's 7000 linguistic varieties the absence is acute. Current approaches cannot hope to meet the resource demand for even a reasonable subset of the languages currently spoken because they seek to document phenomena of great variability using resources such as national funding that are highly constrained in terms of amount, duration and scope. This paper describes efforts to augment the traditional incentives of monetary compensation with alternate incentives in order to elicit greater contributions of linguistic data, metadata and annotation. It also touches on the adjustments to workforces, workflows and post-processing needed to collect and exploit data elicited under novel incentives.

Novel Incentives in Data Collection and Requisite Processing

Saturday 28 May, 14:25 – 15:15
Session Chairperson: Cieri

Herme & Beyond; the Collection of Natural Speech Data

Nick Campbell
This paper describes our approach to the collection of 'natural' (i.e., representative) data from spoken interactions in a social setting in the context of the development (through time) of expressive speech synthesis. Over the past ten years or so, we have collected several corpora of unprompted social conversations that illustrate the 'contact' element of speech that was lacking in many of the corpora collected by use of a specific 'task' with paid participants. The paper discusses the technical and ethical issues of collecting such spoken material, and highlights some of the problems we have encountered in the processing of this much-needed data. Through the use of attractive conversational devices, we have found that natural human curiosity, and an element of social programming combine to provide us with a rich source of material that complements the task-based collections from paid informants.

FKC Corpus: a Japanese Corpus from New Opinion Survey Service

Kensuke Mitsuzawa, Maito Tauchi, Mathieu Domoulin, Masanori Nakashima, Tomoya Mizumoto
In this paper, we present the FKC corpus which is from Fuman Kaitori Center (FKC). The FKC is a Japanese consumer opinion data collection and analysis service. The main advantage of the FKC is the system that awards greater points to user input containing more information, which encourages users to input categorical information. Thanks to this system, the FKC corpus has consumers' opinions with abundant category and user demographics, and is considered to serve multiple NLP tasks: opinion mining, document classification, author inferring and sentiment classification. The FKC corpus consists of 254,683 posts coming from 25,092 users. All posts are checked by annotators who are working for the FKC in crowdsourcing. The posts in the FKC corpus mainly comes from mobile devices, and one third of them are about products or events related to daily life. We also show some correlations between point incentives and users' motivations which keeps them posting their opinions with abundant category information.

The FKC corpus is available under an original license of the FKC. Currently, the FKC gives permission to use directly; thus, those who hope to use the FKC corpus need to send their request to first author.

Novel Incentives and Workflows for Annotation

Saturday 28 May, 15:15 – 16:05
Session Chairperson: Cieri

A Fun and Engaging Interface for Crowdsourcing Named Entities

Kara Greenfield, Kelsey Chan, Joseph P. Campbell
There are many current problems in natural language processing that are best solved by training algorithms on an annotated in-language, in-domain corpus. The more representative the training corpus is of the test data, the better the algorithm will perform, but also the less likely it is that such a corpus has already been annotated. Annotating corpora for natural language processing tasks is typically a time consuming and expensive process. In this paper, we provide a case study in using crowd sourcing to curate an in-domain corpus for named entity recognition, a common problem in natural language processing. In particular, we present our use of fun, engaging user interfaces as a way to entice workers to partake in our crowd sourcing task while avoiding inflating our payments in a way that would attract more mercenary workers than conscientious ones. Additionally, we provide a survey of alternate interfaces for collecting annotations of named entities and compare our approach to those systems.

Novel Incentives for Phrase Detectives

Massimo Poesio, Jon Chamberlain, Udo Kruschwitz and Chris Madge
The Phrase Detectives Game-With-A-Purpose for anaphoric annotation is a moderately successful example of use of novel incentives to create resources for computational linguistics. In this paper we summarize the Phrase Detectives experience in terms of incentives and discuss our future plans to improve such incentives.

Understanding and Exploiting Data from Alternative Sources

Saturday 28 May, 16:30 – 17:20
Session Chairperson: Cieri

Evaluation of Anchor Texts for Automated Link Discovery in Semi-structured Web Documents

Na'im Tyson, Jonathan Roberts, Jeff Allen, Matt Lipson
Using an English noun phrase grammar defined by Hulth (2004a) as a starting point, we created an English noun phrase chunker to extract anchor text candidates identified within web-based articles. These phrases served as candidates for anchor texts linking articles within the About.com network of content sites. Freelance writers—serving as annotators with little to no training outside the domain authority of their respective fields—evaluated articles that received these machine-generated anchor texts using an annotation environment. Unlike other large-scale linguistic annotation projects, where annotators receive an evaluation based on a reference corpus, there was not sufficient time or funding to create a corpus of documents for anchor text comparisons amongst the annotators—thereby complicating the computation of inter-labeler agreement. Instead of using a reference corpus, we assumed that the anchor text generator was another annotator. We then computed the average Cohen's Kappa Coefficient (Landis and Koch, 1977) across all pairings of the anchor text generator and an annotator. Our approach showed a fair agreement level on average (as described in Pustejovsky and Stubbs (2013, p. 131–132)).

Unconventional Approaches to Gathering and Sharing Resources for Spoken Dialog Research

Maxine Eskenazi, Sungjin Lee, Tiancheng Zhao, Ting Yao Hu, Alan W Black

The DialRC and DialPort projects have employed unconventional approaches to data gathering and resource sharing. The projects started sharing by distributing the speech, transcription and logfile data gathered by the Let's Go system. That system has responded to over 220,000 calls from real users of the Allegheny County Port Authority. The Let's Go platform proved to be a very successful way to run studies, with a dataflow of about 1300 dialogs per month. Thus, DialRC built a research platform that was used by other researchers, enabling then to run studies with the Let's Go real users. Challenges were also run on this platform. Finally DialPort follows in the footsteps of DialRC by creating a spoken dialog portal with real users that other dialog systems can be connected to. This paper examines the impact that these activities have had on the spoken dialog research community.

The Future of Incentives, Workforces, Workflows and Data Exploitation

Saturday 28 May, 17:20 – 18:00

Session Chairperson: Cieri

Oral Histories: Linguistic Documentation as Social Media

Mark Liberman

Oral history recordings were pioneered by anthropologists in the early 20th century, collected by Alan Lomax and by the Federal Writers' Project during the 1930s and 1940s, and popularized by authors like Oscar Lewis and Studs Terkel in the 1950s and 1960s. Inexpensive tape recorders allowed the form to spread in the 1960s and 1970s. Now a new opportunity is provided by the combination of ubiquitous multimedia-capable digital devices, inexpensive mass storage, and universally accessible networking. The potential popularity of oral history-like recordings is demonstrated by the tens of thousands of people who have made recordings for StoryCorps. However, there is still no easy way for a motivated group – a family, an athletic group, a school class, a business, a scholarly discipline a club, a church -- to create and publish a collection of oral histories or similar forms of cultural documentation. But the software required to make and edit such recordings, and to transcribe, index, document, publish, and comment on them, is relatively simple and easy to create. And with the right infrastructure, millions of people around the world would participate, creating linguistic and cultural documentation on an unprecedented scale.

Normalisation and Analysis of Social Media Texts (NormSoMe)

28 May 2016

ABSTRACTS

Editors:

Andrius Utka, Jurgita Vaičenonienė, Rita Butkienė

Workshop Programme

28 May, 2016 (morning session)

9:00 – 9:05 – Introduction to the Workshop (by Andrius Utka)

9:05 – 9:45 – Session 1 (Chair: Martin Volk)
Torsten Zesch (keynote speech): *Your noise is my research question! - Limitations of normalizing social media data*

9:45 – 10:35 – Session 2 (Chair: Michi Amsler)
Judit Ács, József Halmi: *Hunaccent: Small Footprint Diacritic Restoration for Social Media*
Andrius Utka, Darius Amilevičius: *Normalisation of Lithuanian Social Media Texts: Towards Morphological Analysis of User-Generated Comments*

10:30 – 11:00 Coffee break

11:00 – 13:00 – Session 3 (Chair: Andrius Utka)
Jaka Čibej, Darja Fišer, Tomaž Erjavec: *Normalisation, Tokenisation and Sentence Segmentation of Slovene Tweets*
Hans van Halteren, Nelleke Oostdijk*: Listening to the Noise: Model Improvement on the Basis of Variation Patterns in Tweets*
Rob van der Goot: *Normalizing Social Media Texts by Combining Word Embeddings and Edit Distances in a Random Forest Regressor*
Ronja Laarmann-Quante, Stefanie Dipper: *An Annotation Scheme for the Comparison of Different Genres of Social Media with a Focus on Normalization*
Tatjana Scheffler, Elina Zarisheva: *Dialog Act Recognition for Twitter Conversations*

Workshop Organizing Committee

Andrius Utka	Vytautas Magnus University, Kaunas
Jolanta Kovalevskaitė	Vytautas Magnus University, Kaunas
Danguolė Kalinauskaitė	Vytautas Magnus University, Kaunas
Martin Volk	University of Zurich
Rita Butkienė	Kaunas University of Technology
Jurgita Vaičenonienė	Vytautas Magnus University, Kaunas

Workshop Programme Committee

Darius Amilevičius	Vytautas Magnus University, Kaunas
Michi Amsler	University of Zurich
Loic Boizou	Vytautas Magnus University, Kaunas
Gintarė Grigonytė	Stockholm University
Jurgita Kapočiūtė-Dzikienė	Vytautas Magnus University, Kaunas
Tomas Krilavičius	Vytautas Magnus University, Kaunas
Joakim Nivre	Uppsala University
Raivis Skadinš	Tilde, Riga, Latvia
Andrius Utka	Vytautas Magnus University, Kaunas
Martin Volk	University of Zurich

Preface

Social media texts provide large quantities of interesting and useful data as well as new challenges for NLP. Social media texts include chats, online commentaries, reviews, blogs, emails, forums, and other genres. Typically, the texts are informal and notoriously noisy. Thus, many NLP tools have difficulties processing and normalizing the data.

As English social media has been investigated most widely, we also invited papers on other languages, especially those rich in inflections and diacritics, which cause additional processing problems. In the programme of NormSoMe, besides English, papers on Dutch, German, Hungarian, Lithuanian, and Slovene are included.

The workshop is aimed at researchers who have solutions, insights, and ideas for tackling the processing of social media texts, or who are interested in this field of research.

Session 2

Date / Time *[Saturday 28 May, 9:45 – 10:35]*
Chairperson: Michi Amsler

Hunaccent: Small Footprint Diacritic Restoration for Social Media

Judit Ács, József Halmi

We present a language-independent method for automatic diacritic restoration. The method focuses on low computational resource usage, making it suitable for mobile devices. We train a decision tree classifier on character-based features without involving a dictionary. Since our features require at most a few characters of context, this approach can be applied to very short text segments such as tweets and text messages. We test the method on a Hungarian web corpus and on Hungarian Facebook comments. It achieves state-of-the-art results on web data and over 92% on Facebook comments. A C++ implementation for Hungarian diacritics is publicly available, support for other languages is under development.

Normalisation of Lithuanian Social Media Texts: Towards Morphological Analysis of User-Generated Comments

Andrius Utka, Darius Amilevičius

In this paper, we present a preliminary research on the normalisation of Lithuanian social media texts. Specifically, the paper deals with language normalisation issues in Lithuanian user-generated comments in the three popular websites: Lietuvos Rytas (Lithuanian Morning), Verslo žinios (Business News), and Delfi.lt. We have established the proportion of out-of-vocabulary (OOV) words in the dataset by using a standard Lithuanian tokenizer and a morphological analyser from the newly developed Information System for Semantic and Syntactic Analysis of the Lithuanian Language (LKKSAIS). A detailed qualitative analysis of extracted OOV words is presented, where specific aspects of Lithuanian social media texts are determined: namely, the extent of missing diacritics, as well as other prevalent error types. A standard Lithuanian spell checker is used for the restoration of missing diacritics and correction of other errors in user-generated comments, which considerably improves the morphological analysis.

Session 3

Date / Time *[Saturday 28 May, 11:00 – 13:00]*
Chairperson: Andrius Utka

Normalisation, Tokenisation and Sentence Segmentation of Slovene Tweets

Jaka Čibej, Darja Fišer, Tomaž Erjavec

Online user-generated content such as posts on social media, blogs, and forums, is becoming an increasingly important source of information, as shown by numerous rapidly growing NLP fields such as sentiment analysis and data mining. However, user-generated content is well-known to contain a significant degree of noise, e.g. abbreviations, missing spaces, as well as non-standard spelling, lexis, and use of punctuation. All this hinders the effectiveness of NLP tools when processing such data, and to overcome this obstacle, data normalisation is required. In this paper, we present a training set that will be used to improve the tokenisation, normalisation, and sentence segmentation of Slovene tweets. We describe some of the most Twitter-specific aspects of our annotation guidelines as well as the workflow of our annotation campaign, the goal of which was to create a manually annotated gold-standard dataset of 4,000 tweets extracted from the JANES corpus of Internet Slovene.

Listening to the Noise: Model Improvement on the Basis of Variation Patterns in Tweets

Hans van Halteren, Nelleke Oostdijk

In this paper, we take the view that the wide diversity in the language (use) found on Twitter can be explained by the fact that language use varies between users and from one use situation to another: what users are tweeting about and to what audience will influence the choices users make. We propose to model the language use of Twitter tribes, i.e. peer groups of users tweeting in different use situations. We argue that the use of tribal models can improve the modeling of the substantial variation present in Twitter (and other social media), and that the resulting models can be used in the normalization of text for NLP tasks. In our discussion of variation at the linguistic levels of orthography, spelling, and syntax, we give numerous examples of various types of variation, and indicate how tribal models could help process text in which such variation occurs. All examples are derived from our own experience with the Dutch part of Twitter, for which we could draw on a multi-billion word dataset.

Normalizing Social Media Texts by Combining Word Embeddings and Edit Distances in a Random Forest Regressor

Rob van der Goot

In this work, we adapt the traditional framework for spelling correction to the more novel task of normalization of social media content. To generate possible normalization candidates, we complement the traditional approach with a word embeddings model. To rank the candidates we will use a random forest regressor, combining the features from the generation with some N-gram features. The N-gram model contributes significantly to the model, because no other features account for short-distance relations between words. A random forest regressor fits this task very well, presumably because it can model the different types of corrections. Additionally we show that 500 annotated sentences should be enough training data to train this system reasonably well on a new domain. Our proposed system performs slightly worse compared to the state-of-the-art. The main advantage is the simplicity of the model, allowing for easy expansions.

An Annotation Scheme for the Comparison of Different Genres of Social Media with a Focus on Normalization

Ronja Laarmann-Quante, Stefanie Dipper

Most work on automatic normalization of social media data is restricted to a specific communication medium and often guided by noncomprehensive notions of which phenomena have to be normalized. This paper aims to shed light on the questions (a) what kinds of deviations from the standard' can be found in German social media and (b) how these differ across different genres of computer-mediated communication (CMC). To study these issues systematically, we propose a comprehensive annotation scheme which categorizes 'nonstandard' (defined as out-of-vocabulary, OOV) tokens of various genres of CMC with a focus on the challenges they pose to automatic normalization. In a pilot study, we achieved a high inter-annotator agreement (Cohen's κ > .8), which suggests good applicability of the scheme. Primary results indicate that the predominant phenomena are rather diverse across genres and, furthermore, that in some genres, general OOV-tokens, which are not CMC-specific (like named entities or regular non-listed words), play a more dominant role than one might guess at first sight.

Dialog Act Recognition for Twitter Conversations

Tatjana Scheffler, Elina Zarisheva

In this paper, we present our approach to dialog act classification for German Twitter conversations. In contrast to previous work, we took the entire conversation context into account and classified individual segments within tweets (a tweet can contain more than one segment). In addition, we used fine-grained dialog act annotations with a taxonomy of 56 categories. We trained three classifiers with different feature sets. The best results are achieved with CRF sequence taggers. For the full DA taxonomy, we achieved an f-measure of up to 0.31, for the reduced taxonomy (12 DAs), up to 0.51, and minimal taxonomy (8 DAs), 0.72, showing that dialog act recognition on Twitter conversations is quite reliable for small taxonomies. The results improve on previous work on speech act classification for social media posts. The improvement is due to two factors: (i) Our classifiers explicitly model the sequential structure of conversations, whereas previous approaches classify individual social media posts without taking dialog structure into account. (ii) We segmented tweets into utterances first (segmentation is not a part of this work), while all previous approaches assign exactly one speech act to a post. In our corpus, over 30% of tweets consist of several speech acts.

4th Workshop on Challenges in the Management of Large Corpora

28 April 2016

ABSTRACTS

Editors:

Piotr Bański, Adrien Barbaresi, Hanno Biber, Evelyn Breiteneder, Simon Clematide, Marc Kupietz, Harald Lüngen, Andreas Witt

Workshop Programme

28 May 2016

14:00-16:00 – Session A

Introduction

Jochen Tiepmar,
CTS Text Miner – Text Mining Framework based on the Canonical Text Services Protocol

Jelke Bloem,
Evaluating Automatically Annotated Treebanks for Linguistic Research

Marcin Junczys-Dowmunt, Bruno Pouliquen and Christophe Mazenc,
COPPA V2.0: Corpus of Parallel Patent Applications. Building Large Parallel Corpora with GNU Make

Johannes Graën, Simon Clematide and Martin Volk,
Efficient Exploration of Translation Variants in Large Multiparallel Corpora using a Relational Database

16:00-16:30 Coffee break

16:30-18:00 – Session B

Adrien Barbaresi,
Collection and Indexation of Tweets with a Geographical Focus

Ruxandra Cosma, Dan Cristea, Marc Kupietz, Dan Tufiș and Andreas Witt,
DRuKoLA – Towards Contrastive German-Romanian Research based on Comparable Corpora

Svetla Koeva, Ivelina Stoyanova, Maria Todorova, Svetlozara Leseva and Tsvetana Dimitrova,
Metadata Extraction, Representation and Management within the Bulgarian National Corpus

Closing Remarks

Editors/Workshop Organizers

Piotr Bański, Marc Kupietz, Harald Lüngen, Andreas Witt	Institut für Deutsche Sprache, Mannheim
Adrien Barbaresi, Hanno Biber, Evelyn Breiteneder	Institute for Corpus Linguistics and Text Technology, Vienna
Simon Clematide	Institute of Computational Linguistics, Zurich

Workshop Programme Committee

Steve Cassidy	Macquarie University
Damir Ćavar	Indiana University, Bloomington
Isabella Chiari	Sapienza University of Rome
Dan Cristea	"Alexandru Ioan Cuza" University of Iaşi
Václav Cvrček	Charles University Prague
Koenraad De Smedt	University of Bergen
Tomaž Erjavec	Jožef Stefan Institute
Andrew Hardie	Lancaster University
Serge Heiden	ENS de Lyon
Nancy Ide	Vassar College
Miloš Jakubíček	Lexical Computing Ltd.
Piotr Pęzik	University of Łódź
Uwe Quasthoff	Leipzig University
Paul Rayson	Lancaster University
Laurent Romary	INRIA, DARIAH
Roland Schäfer	FU Berlin
Serge Sharoff	University of Leeds
Marko Tadić	University of Zagreb, Faculty of Humanities and Social Sciences
Ludovic Tanguy	University of Toulouse
Dan Tufiş	Romanian Academy, Bucharest
Tamás Váradi	Research Institute for Linguistics, Hungarian Academy of Sciences

Workshop Homepage

http://corpora.ids-mannheim.de/cmlc-2016.html

Preface/Introduction

Creating very large corpora no longer appears to be a challenge. With the constantly growing amount of born-digital text – be it available on the web or only on the servers of publishing companies – and with the rising number of printed texts digitized by public institutions or technological giants such as Google, we may safely expect the upper limits of text collections to keep increasing for years to come. Although some of this was already true 20 years ago, we have a strong impression that the challenge has now shifted from an increase in terms of size to the effective and efficient processing of the large amounts of primary data and much larger amounts of annotation data.

On the one hand, some fundamental technical methods and strategies call for re-evaluation. These include, for example, efficient and sustainable curation of data, management of collections that span multiple volumes or that are distributed across several centres, innovative corpus architectures that maximize the usefulness of data, and techniques that allow for efficient search and analysis.

On the other hand, the new challenges require research into language-modelling methods and new corpus-linguistic methodologies that can make use of extremely large, structured datasets. These methodologies must re-address the tasks of investigating rare phenomena involving multiple lexical items, of finding and representing fine-grained sub-regularities, and of investigating variations within and across language domains. This should be accompanied by new methods to structure both content and search results, in order to, among others, cope with false positives, assess data quality, or ensure interoperability. Another much-needed research goal is visualization techniques that facilitate the interpretation of results and formulation of new hypotheses.

Due to the interest that the first meeting (at LREC 2012 in Istanbul) of CMLC enjoyed, the workshop has become a cyclic event. The second meeting took place at LREC again, in 2014 in Reykjavík; the third edition of CMLC was part of Corpus Linguistics 2015 in Lancaster. The coming fourth meeting will take place in Portorož, Slovenia, as part of LREC-2016.

Session A

Saturday, 28th May, 14:00 – 16:00

CTS Text Miner -Text Mining Framework based on the Canonical Text Services Protocol

Jochen Tiepmar

The purpose of this paper is to describe a modular framework for text mining that uses Canonical Text Service (CTS) as a data source. By combining standardized functionalities with standardized access to text data, this framework intends to reduce the heterogeneity of workflows in today's Digital Humanities and act as an important element of a text research infrastructure.

For this work the implementation of the CTS protocol described in (Tiepmar, 2015) is used. It uses advanced functionalities that are not part of the specifications of CTS. This means that, while most current modules should work with different implementations of the CTS protocol, it cannot be guaranteed that any future module will work.

Evaluating Automatically Annotated Treebanks for Linguistic Research

Jelke Bloem

This study discusses evaluation methods for linguists to use when employing an automatically annotated treebank as a source of linguistic evidence. While treebanks are usually evaluated with a general measure over all the data, linguistic studies often focus on a particular construction or a group of structures. To judge the quality of linguistic evidence in this case, it would be beneficial to estimate annotation quality over all instances of a particular construction. I discuss the relative advantages and disadvantages of four approaches to this type of evaluation: manual evaluation of the results, manual evaluation of the text, falling back to simpler annotation and searching for particular instances of the construction. Furthermore, I illustrate the approaches using an example from Dutch linguistics, two-verb cluster constructions, and estimate precision and recall for this construction on a large automatically annotated treebank of Dutch. From this, I conclude that a combination of approaches on samples from the treebank can be used to estimate the accuracy of the annotation for the construction of interest. This allows researchers to make more definite linguistic claims on the basis of data from automatically annotated treebanks.

COPPA V2.0: Corpus of Parallel Patent Applications. Building Large Parallel Corpora with GNU Make

Marcin Junczys-Dowmunt, Bruno Pouliquen and Christophe Mazenc

WIPO seeks to help users and researchers to overcome the language barrier when searching patents published in different languages. Having collected a big multilingual corpus of translated patent applications, WIPO decided to share this corpus in a product called COPPA (Corpus Of Parallel Patent Applications) to stimulate research in Machine translation and in language tools for patent texts.

A first version was released in 2011 but contained only French and English languages. It has been decided to release a major update of this product containing newer data (from 2011 up to 2014) but also other languages (German, English, French, Japanese, Korean, Portuguese, Spanish, Russian and Chinese). This corpus can be used for terminology extraction, cross-language information retrieval or statistical machine translation. With the new version a huge number of files (more than 26 million) has to be processed. We describe the technical process in details.

Efficient Exploration of Translation Variants in Large Multiparallel Corpora using a Relational Database

Johannes Graën, Simon Clematide and Martin Volk

We present an approach for searching and exploring translation variants of multi-word units in large multiparallel corpora based on a relational database management system. Our web-based application Multilingwis, which allows for multilingual lookups of phrases and words in English, French, German, Italian and Spanish, is of interest to anybody who wants to quickly compare expressions across several languages, such as language learners without linguistic knowledge. In this paper, we focus on the technical aspects of how to represent and efficiently retrieve all occurrences that match the user's query in one of five languages simultaneously with their translations into the other four languages. In order to identify such translations in our corpus of 220 million tokens in total, we use statistical sentence and word alignment. By using materialized views, composite indexes, and pre-planned search functions, our relational database management system handles large result sets with only moderate requirements to the underlying hardware. As our systematic evaluation on 200 search terms per language shows, we can achieve retrieval times below 1 second in 75 % of the cases for multi-word expressions.

Session B

Saturday, 28th May, 16:30 – 18:00

Collection and Indexation of Tweets with a Geographical Focus

Adrien Barbaresi

This paper introduces a Twitter corpus currently focused geographically in order to (1) test selection and collection processes for a given region and (2) find a suitable database to query, filter, and visualize the tweets. Due to access restrictions, it is not possible to retrieve all available tweets, which is why corpus construction implies a series of decisions described below. The corpus focuses on Austrian users, as data collection grounds on a two-tier detection process addressing corpus construction and user location issues. The emphasis lies on short messages whose sender mentions a place in Austria as his/her hometown or tweets from places located in Austria. The resulting user base is then queried and enlarged using focused crawling and random sampling, so that the corpus is refined and completed in the way of a monitor corpus. Its current volume is 21.7 million tweets from approximately 125,000 users. The tweets are indexed using Elasticsearch and queried via the Kibana frontend, which allows for queries on metadata as well as for the visualization of geolocalized tweets (currently about 3.3% of the collection).

DRuKoLA – Towards Contrastive German-Romanian Research based on Comparable Corpora

Ruxandra Cosma, Dan Cristea, Marc Kupietz, Dan Tufiș and Andreas Witt

This paper introduces the recently started DRuKoLA-project that aims at providing mechanisms to flexibly draw virtual comparable corpora from the German Reference Corpus DeReKo and the Reference Corpus of Contemporary Romanian Language CoRoLa in order to use these virtual corpora as empirical basis for contrastive linguistic research.

Metadata Extraction, Representation and Management within the Bulgarian National Corpus

Svetla Koeva, Ivelina Stoyanova, Maria Todorova, Svetlozara Leseva and Tsvetana Dimitrova

This paper presents the extraction, representation and management of metadata in the Bulgarian National Corpus. We briefly present the current state of the Corpus and the general principles on which its development lies: uniformity, diversity of text samples, automatic compilation, extensive metadata, multi-layered linguistic annotation. The relevant information for the texts in the Corpus is stored into different types of metadata categories: administrative, editorial, structural, descriptive, classificational, analytical, and statistical metadata. The structure and the design of the Bulgarian National Corpus is flexible and can incorporate new metadata categories and values.

Further, we discuss some of the automatic procedures for extraction of metadata applied in the compilation of the Bulgarian National Corpus: (i) metatextual techniques – extracting information from the HTML/XML markup of the original files through a combination of automatic and manual procedures; and (ii) textual techniques – applying text analysis and heuristics using a set of language resources. We briefly present the MetadataEditor – a tool for manual metadata editing and verification. Directions for future work on the extraction, representation and management of metadata include development of more advanced techniques for language processing, domain-specific analysis, and verification procedures.

Just Talking – Casual Talk among Humans and Machines

28 May 2016

ABSTRACTS

Editors:

Emer Gilmartin, Nick Campbell

Workshop Programme

28 May 2016

09:00 – 09:10 – Introduction by Workshop Chair

09:10 – 10:30 – Paper Session 1
Jens Allwood and Elisabeth Ahlsen, *Small talk and its role in different social activities - a corpus based analysis*
Stefan Olafsson and Timothy Bickmore, *"That reminds me...": Towards a Computational Model of Topic Development Within and Across Conversations*
Hanae Koiso, Yayoi Tanaka, Ryoko Watanabe and Yasuharu Den, *A Large-Scale Corpus of Everyday Japanese Conversation: On Methodology for Recording Naturally Occurring Conversations*
Saturnino Luz, Nick Campbell and Fasih Haider, *Title of the paper*

10:30 – 11:00 Coffee break

11:00 – 12:00 – Paper Session 2
Katri Hiovain and Kristiina Jokinen, *Different Types of Laughter in North Sami Conversational Speech*
Kevin El Haddad, Huseyin Cakmak, Stéphane Dupont and Thierry Dutoit, *Laughter and Smile Processing for Human-Computer Interactions*
Emer Gilmartin, Ketong Su, Yuyun Huang, Kevin El Haddad, Christy Elias, Benjamin R. Cowan and Nick Campbell, *Making Idle Talk: Designing and Implementing Casual Talk*

16:00 – 16:30 Coffee break

12:00 – 12:45 – Discussion Session – Getting Natural Talk

Workshop Organizers

Nick Campbell	Trinity College, Dublin
Emer Gilmartin	Trinity College, Dublin
Laurence Devillers	LIMSI, Paris
Sophie Rosset	LIMSI, Paris
Guillaume Dubuisson Duplessis	LIMSI, Paris

Workshop Programme Committee

Nick Campbell	Trinity College, Dublin
Emer Gilmartin	Trinity College, Dublin
Laurence Devillers	LIMSI, Paris
Sophie Rosset	LIMSI, Paris
Guillaume Dubuisson Duplessis	LIMSI, Paris

Introduction

This workshop focusses on the collection and analysis of resources, novel research, and applications in both human-human and human-machine casual interaction. A major distinction between different types of spoken interaction is whether the goal is ‘transactional’ or ‘interactional’. Transactional, or task-based, talk has short-term goals which are clearly defined and known to the participants – as in service encounters in shops or business meetings. Task-based conversations rely heavily on the transfer of linguistic or lexical information. In technology, most spoken dialogue systems have been task-based for reasons of tractability, concentrating on practical activities such as travel planning. However, in real-life social talk there is often no obvious short term task to be accomplished through speech and the purpose of the interaction is better described as building and maintaining social bonds and transferring attitudinal or affective information – examples of this interactional talk include greetings, gossip, and social chat or small talk. A tenant’s short chat about the weather with the concierge of an apartment block is not intended to transfer important meteorological data but rather to build a relationship which may serve either of the participants in the future. Of course, most transactional encounters are peppered with social or interactional elements as the establishment and maintenance of friendly relationships contributes to task success. There is increasing interest in modelling interactional talk for applications including social robotics, education, health and companionship. In order to successfully design and implement these applications, there is a need for greater understanding of the mechanics of social talk, particularly its multimodal features. This understanding relies on relevant language resources (corpora, analysis tools), analysis, and experimental technologies. This workshop provides a focal point for the growing research community on social talk to discuss available resources and ongoing work.

Paper Session 1

Saturday 28 May, 9:10 – 10:30

Small talk and its role in different social activities - a corpus based analysis

Jens Allwood and Elisabeth Ahlsen

This study investigates the occurrence and role of small talk in a number of different social activities, based on video-recorded corpus data from the GSLC (The Gothenburg Spoken Language Corpus) which represents a broad range of different social activities. The study builds on findings from studying communication in different social activity types and compares them with respect to the occurrence, content and role of small talk. The purpose is (i) to describe the characteristics of small talk in general and (ii) to investigate whether and, in that case, in what respects the nature of small talk varies depending on the social activity where it occurs. General types of small talk are found, which can occur in most activity types, for example talk about the weather, the family or the activity at hand. Other types of small talk depend on activity specific factors, such as how formal or informal the activity is, the background and activity roles of the participants, factors in the environment and typical interaction patterns for the activity.

"That reminds me...": Towards a Computational Model of Topic Development Within and Across Conversations

Stefan Olafsson and Timothy Bickmore

We aim to increase user engagement with dialog systems over long periods of time by developing a computational model of longitudinal topic development. Examining a corpus of interactional talk made it clear that shifts and changes in topic are not random and have different manifestations depending on how long the participants have known each other. We developed and evaluated an annotation scheme based on interactional dialog theories that will inform the creation of a computational model of topic development across conversations.

A Large-Scale Corpus of Everyday Japanese Conversation: On Methodology for Recording Naturally Occurring Conversations

Hanae Koiso, Yayoi Tanaka, Ryoko Watanabe and Yasuharu Den

In 2016, we set about building a large-scale corpus of everyday Japanese conversation?a collection of conversations embedded in naturally occurring activities in daily life. We will collect more than 200 hours of recordings over six years, publishing the corpus in 2022. Before building such a corpus, we have conducted a pilot project, whose purposes are i) to establish a corpus design for collecting various kinds of everyday conversations in a balanced manner, ii) to develop a methodology for recording naturally occurring conversations, and iii) to create a transcription system suitable for precisely and efficiently transcribing natural conversations. This paper focuses on the second issue. We first describe two types of methods for recording various kinds of conversations embedded in naturally occurring activities in everyday situations. Next, we show recording devices we use and sample data. Finally, we discuss ethical and other issues, including the collection of consent forms and questionnaires.

Data Collection Using a Real Time Feedback Tool for Non Verbal Presentation Skills Training

Saturnino Luz, Nick Campbell and Fasih Haider

This paper describes the data recording setting and systems used in the first pilot data collection activity for the METALOGUE project. The objective of this activity is to provide an opportunity for participants to use a ‘real time presentation trainer’ feedback tool during debate. In total 2 sessions have been recorded and the data is made available to all participants of the project.

Paper Session 2

Saturday 28 May, 11:00 – 12:00

Different Types of Laughter in North Sami Conversational Speech

Katri Hiovain and Kristiina Jokinen

This paper describes how various laughter types differ from each other acoustically in a North Sami conversational speech corpus collected and annotated within the DigiSami project. The laughter annotations were done with Praat and included two tag types, the first of which indicated if the laugh was a free laugh or speech-talk (laughing speech), and the second one indicating more specific laughter type. In our study, pitch, duration and intensity were extracted for laughter bouts representing every laughter type, and the paper describes the first analysis of the data. The conversational speech of the Sami languages has not yet been systematically studied, so our analysis can be compared with the results of laughter studies conducted in other languages, while also contributing to empirical observations of the North Sami language.

Laughter and Smile Processing for Human-Computer Interactions

Kevin El Haddad, Huseyin Cakmak, Stéphane Dupont and Thierry Dutoit

This paper provides a short summary of the importance of taking into account laughter and smile expressions in Human-Computer Interaction systems. Based on the literature, we mention some important characteristics of these expressions in our daily social interactions. We describe some of our own contributions and ongoing work to this field.

Making Idle Talk: Designing and Implementing Casual Talk

Emer Gilmartin, Ketong Su, Yuyun Huang, Kevin El Haddad, Christy Elias, Benjamin R. Cowan and Nick Campbell

While casual conversation or talk is ubiquitous in human life, it has not been taxonomised or indeed recreated in spoken dialogue applications to the extent that more tractable task-based or practical dialogues have been. With the recent surge in interest in more sociable speaking systems for use in companionship, educational, and healthcare applications, there is increasing need for clearer understanding at several levels of what casual talk is and how it may be modelled. In this paper we describe our current work on exploring, creating, and evaluating human-machine casual social talk.

RE-WOCHAT: Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents - Development and Evaluation

May 28th, 2016

ABSTRACTS

Editors:

Rafael E. Banchs, Ryuichiro Higashinaka, Wolfgang Minker, Joseph Mariani, David Traum

Workshop Programme

13:50 – 14:00 – Welcome Message by the Organizing Team

14:00 – 15:00 – Short Paper Session I

Data Collection for Interactive Learning through the Dialog
Miroslav Vodolán, Filip Jurčíček

A Context-aware Natural Language Generation Dataset for Dialogue Systems
Ondřej Dušek, Filip Jurčíček

Dead Man Tweeting
David Nilsson, Magnus Sahlgren, Jussi Karlgren

15:00 – 16:00 – Short Paper Session II

Chatbot Evaluation and Database Expansion via Crowdsourcing
Zhou Yu, Ziyu Xu, Alan W. Black, Alexander I. Rudnicky

Framework for the Formulation of Metrics for Conversational Agent Evaluation
Mohammed Kaleem, Omar Alobadi, James O'Shea, Keeley Crockett

Tracking Conversational Gestures of Extraverts and Introverts in Multimodal Interaction
David Novick, Adriana Camacho, Ivan Gris, Laura M. Rodriguez

16:00 – 16:30 – Coffee break

16:30 – 17:00 – Long Paper Session

Automatic Extraction of Chatbot Training Data from Natural Dialogue Corpora
Bayan Abu Shawar, Eric Atwell

17:00 – 17:50 – Shared Task Presentation and Poster Session

Shared Task on Data Collection and Annotation
Luis Fernando D'Haro, Bayan Abu Shawar, Zhou Yu

Shared Task Chatbot Description Reports:

POLITICIAN, David Kuboň, Barbora Hladká

JOKER, Guillaume Dubuisson Duplessis, Vincent Letard, Anne-Laure Ligozat, Sophie Rosset

IRIS – Informal Response Interactive System, Rafael E. Banchs, Haizhou Li

PY-ELIZA – A Python-based implementation of Eliza, Luis Fernando D'Haro

TICK TOCK, Zhou Yu, Ziyu Xu, Alan W Black, Alexander I. Rudnicky

SARAH Chatbot, Bayan AbuShawar

17:50 – 18:00 – Concluding Remarks by the Organizing Team

Workshop Organizers

Rafael E. Banchs	Institute for Infocomm Research, Singapore
Ryuichiro Higashinaka	Nippon Telegraph and Telephone Corporation, Japan
Wolfgang Minker	Ulm University, Germany
Joseph Mariani	IMMI & LIMSI-CNRS, France
David Traum	University of Southern California, USA

Shared Task Co-organizers

Bayan Abu Shawar	Arab Open University, Jordan
Luis Fernando D'Haro	Agency for Science, Technology and Research, Singapore
Zhou Yu	Carnegie Mellon University, USA

Workshop Programme Committee

Björn Schuller	Imperial College London, UK
David Suendermann	Educational Testing Service (ETS), USA
Diane Litman	University of Pittsburgh, USA
Dilek Hakkani-Tur	Microsoft Research, USA
Gabriel Skantze	KTH Royal Institute of Technology, Sweden
Haizhou Li	Institute for Infocomm Research, Singapore
Jason Williams	Microsoft Research, USA
Jiang Ridong	Agency for Science, Technology and Research, Singapore
Justine Cassell	Carnegie Mellon University, USA
Kristiina Jokinen	University of Helsinki, Finland
Kotaro Funakoshi	Honda Research Institute, Japan
Laurence Devillers	LIMSI-CNRS, France
Luisa Coheur	Lisbon University, Portugal
Matthew Henderson	Google Research, USA
Michael McTear	University of Ulster, UK
Mikio Nakano	Honda Research Institute, Japan
Nick Campbell	Trinity College Dublin, Ireland
Oliver Lemon	Heriot-Watt University, UK
Ramón López-Cózar	University of Granada, Spain
Sakriani Sakti	Nara Institute of Science and Technology, Japan
Satoshi Nakamura	Nara Institute of Science and Technology, Japan
Seokhwan Kim	Agency for Science, Technology and Research, Singapore
Sophie Rosset	LIMSI-CNRS, France
Stefan Ultes	Cambridge University, UK
Suraj Nair	Technische Universität München, Germany
Teruhisa Misu	Honda Research Institute, USA
Tomoki Toda	Nara Institute of Science and Technology, Japan
Yasuharo Den	Chiba University, Japan
Shiv Vitaladevuni	Amazon, USA

Introduction to RE-WOCHAT

Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents – Development and Evaluation

Although non-goal-oriented dialogue systems have been around for many years (more than forty years indeed, if we consider Weizenbaum's Eliza system as the starting milestone), they have been recently gaining a lot of popularity in both research and commercial arenas. From the commercial stand point, non-goal-oriented dialogue seems to be providing an excellent means to engage users for entertainment purposes, as well as to give a more human-like appearance to established vertical goal-driven dialogue systems.

From the research perspective, on the other hand, this kind of systems poses interesting challenges and problems to the research community. Different from goal-oriented dialogue systems, which are domain focused, non-goal-oriented or chat-oriented dialogue requires dealing with knowledge on a wide diversity of domains as well as common sense knowledge related to daily life experiences. Additionally, due to the lack of specific goals in chat-oriented dialogue, this kind of systems cannot be objectively evaluated by using goal completion rates, as in the case of goal-oriented dialogue engines. Moreover, many task-oriented dialogue systems use length of the dialogue as a metric, with a penalty in reward function or assumed user-satisfaction for longer dialogues. For chat dialogue, however, this metric is often reversed: the more interested and enjoyable the chat dialogue is, the longer users will talk with a system.

In this regards, the RE-WOCHAT initiative aims at providing and consolidating a venue for the research community to explore and discuss the state-of-the-art in non-goal-oriented dialogue and its related problems, including resource generation and evaluation. The workshop also accommodates a shared task on "Data Collection and Annotation" aiming at developing and testing a new evaluation framework for non-goal-oriented dialogue engines.

RE-WOCHAT is the result of a working committee initiative on "Automatic Evaluation and Resources" generated during the Shonan Meeting "The Future of Human-Robot Spoken Dialogue: from Information Services to Virtual Assistants" held in Shonan, Japan, at the end of March 2015. The main objective of this meeting was to discuss about the most relevant and promising future directions of research in dialogue systems. The discussion was centred on how these directions should address the different problems and limitations of current dialogue systems, as well as how they can provide the basis for the next generation of intelligent artificial agents. (More information is available at http://shonan.nii.ac.jp/shonan/wp-content/uploads/2011/09/No.2015-7.pdf).

The workshop also constitutes the natural extension of two successful Special Sessions on "Chatbots and Dialogue Agents" collocated with APSIPA conferences in 2014 and 2015 (http://www.apsipa2014.org/home/program/special-sessions and http://www.apsipa2015.org/), and two evaluation workshops related to chat-oriented dialogue systems: dialogue breakdown detection challenge (https://sites.google.com/site/dialoguebreakdowndetection/) and NTCIR short text conversation (http://ntcir12.noahlab.com.hk/stc.htm).

Rafael E. Banchs, Ryuichiro Higashinaka, Wolfgang Minker, Joseph Mariani, David Traum
Portorož, Slovenia, May 28th, 2016

Short Paper Session I

Saturday 28 May, 14:00 – 15:00
Chairperson: Joseph Mariani

Data Collection for Interactive Learning through the Dialog

Miroslav Vodolán, Filip Jurčíček

This paper presents a dataset collected from natural dialogs which enables to test the ability of dialog systems to learn new facts from user utterances throughout the dialog. This interactive learning will help with one of the most prevailing problems of open domain dialog system, which is the sparsity of facts a dialog system can reason about. The proposed dataset, consisting of 1900 collected dialogs, allows simulation of an interactive gaining of denotations and questions explanations from users which can be used for the interactive learning.

A Context-aware Natural Language Generation Dataset for Dialogue Systems

Ondřej Dušek, Filip Jurčíček

We present a novel dataset for natural language generation (NLG) in spoken dialogue systems which includes preceding context (user utterance) along with each system response to be generated, i.e., each pair of source meaning representation and target natural language paraphrase. We expect this to allow an NLG system to adapt (entrain) to the user's way of speaking, thus creating more natural and potentially more successful responses. The dataset has been collected using crowdsourcing, with several stages to obtain natural user utterances and corresponding relevant, natural, and contextually bound system responses. The dataset is available for download under the Creative Commons 4.0 BY-SA license.

Dead Man Tweeting

David Nilsson, Magnus Sahlgren, Jussi Karlgren

This paper presents a prototype — Dead Man Tweeting — of a system that learns semantic avatars from (dead) people's texts, and makes the avatars come alive on Twitter. The system includes a language model for generating sequences of words, a topic model for ensuring that the sequences are topically coherent, and a semantic model that ensures the avatars can be productive and generate novel sequences. The avatars are connected to Twitter and are triggered by keywords that are significant for each particular avatar.

Short Paper Session II

Saturday 28 May, 15:00 – 16:00
Chairperson: Wolfgang Minker

Chatbot Evaluation and Database Expansion via Crowdsourcing

Zhou Yu, Ziyu Xu, Alan W. Black, Alexander I. Rudnicky

Chatbots use a database of responses often culled from a corpus of text generated for a different purpose, for example film scripts or interviews. One consequence of this approach is a mismatch between the data and the inputs generated by participants. We describe an approach that while starting from an existing corpus (of interviews) makes use of crowdsourced data to augment the response database, focusing on responses that people judge as inappropriate. The long term goal is to create a data set of more appropriate chat responses; the short term consequence appears to be the identification and replacement of particularly inappropriate responses. We found the version with the expanded database was rated significantly better in terms of the response level appropriateness and the overall ability to engage users. We also describe strategies we developed that target certain breakdowns discovered during data collection. Both the source code of the chatbot, TickTock, and the data collected are publicly available.

Framework for the Formulation of Metrics for Conversational Agent Evaluation
Mohammed Kaleem, Omar Alobadi, James O'Shea, Keeley Crockett
The evaluation of conversational agents is an area that has not seen much progress since the initial developments during the early 90's. The initial challenge faced when evaluating conversational agents is trying to formulate the metrics that need to be captured and measured in order to gauge the success of a particular conversational agent. Although frameworks exist they overlook the individual objectives of modern conversational agents which are much more than just question answering systems. This paper presents a new framework that has been utilised to formulate metrics to evaluate two conversational agents deployed in two significantly different contexts.

Tracking Conversational Gestures of Extraverts and Introverts in Multimodal Interaction
David Novick, Adriana Camacho, Ivan Gris, Laura M. Rodriguez
Much of our current research explores differences between extraverts and introverts in their perception and production of gestures in multimodal interaction with embodied conversational agent (ECA). While several excellent corpora of conversational gestures have been collected, these corpora do not distinguish conversants by personality dimensions. To enable study of the differences in distribution of conversational gestures between extraverts and introverts, we tracked and automatically annotated gestures of 59 subjects interacting with an ECA in the "Survival on Jungle Island" immersive multimodal adventure. Our work in developing provides an initial corpus for analysis of gesture differences, based on Jung's four personality-type dimensions, of humans interacting with an ECA and suggests that it may be feasible to annotate gestures automatically in real time, based on a gesture lexicon. Preliminary analysis of the automated annotations suggests that introverts more frequently performed gestures in the gesture lexicon than did extraverts.

Long Paper Session

Saturday 28 May, 16:30 – 17:00
Chairperson: Rafael E. Banchs

Automatic Extraction of Chatbot Training Data from Natural Dialogue Corpora
Bayan Abu Shawar, Eric Atwell
A chatbot is a conversational agent that interacts with the users turn by turn using natural language. Different chatbots or human-computer dialogue systems have been developed using spoken or text communication and have been applied in different domains such as: linguistic research, language education, customer service, web site help, and for fun. However, most chatbots are restricted to knowledge that is manually "hand coded" in their files, and to a specific natural language which is written or spoken. This paper presents the program we developed to convert a machine readable text (corpus) to a specific chatbot format, which is then used to retrain a chatbot and generate a chat which is closer to human language. Different corpora were used: dialogue corpora such as the British National Corpus of English (BNC); the holy book of Islam Qur'an which is a monologue corpus where verse and following verse are turns; and the FAQ where questions and answers are pair of turns. The main goal of this automation process is the ability to generate different chatbot prototypes that spoke different languages based on corpus.

Shared Task Presentation and Poster Session

Saturday 28 May, 17:00 – 17:50
Chairperson: Ryuichiro Higashinaka

Shared Task on Data Collection and Annotation

Luis Fernando D'Haro, Bayan AbuShawar, Zhou Yu

This report presents and describes the shared task on “Data Collection and Annotation” conducted in RE-WOCHAT. We describe the main road map envisaged for this and future shared tasks, as well as the proposed collection and annotation schemes. We also summarize the result of the shared task in terms of chatbot platforms made available for it and the amount of collected chatting sessions and annotations.

Politician

David Kuboň, Barbora Hladká

We present a question-answering system Politician designed as a chatbot imitating a politician. It answers questions on political issues. The questions are analyzed using natural language processing techniques and no complex knowledge base is involved. The language used for the interviews is Czech.

Joker Chatterbot

Guillaume Dubuisson Duplessis, Vincent Letard, Anne-Laure Ligozat, Sophie Rosset

The Joker chatterbot is an example-based system that uses a database of indexed dialogue examples automatically built from a television drama subtitle corpus to manage social open-domain dialogue.

IRIS (Informal Response Interactive System)

Rafael E. Banchs, Haizhou Li

This report describes IRIS (Informal Response Interactive System), a chat-oriented dialogue system based on the vector space model framework. IRIS was one of the systems made available as part of the RE-WOCHAT Shared Task platform for collecting human-chatbot dialogue sessions.

Py-Eliza: A Python-based Implementation of the Famous Computer Therapist

Luis Fernando D'Haro

In this report we provide information about the functionalities and capabilities of pyEliza a python-based implementation of the famous Eliza chatbot proposed by Weizenbaum in 1966. PyEliza implements a rule-based chatbot that encourage users to talk about their lives and feelings. In addition to the chat capabilities, this chatbot features some management functionalities to keep track of all interactions with the users by using logs and automatically generating annotated and anonymized XML files.

TickTock

Zhou Yu, Ziyu Xu, Alan W Black, Alexander I. Rudnicky

This is a description of the TickTock chatbot system, which is a retrieval based system that utilizes conversational strategies to improve the system performance. It has two versions, one with multimodal signals as input; one with text input through typing. The multimodal version is a stand-alone system, while the text version is a web-API version. In this report, we focus on describing the web-API version of TickTock, which is used in the shared task.

Sarah Chatbot

Bayan AbuShawar

Sarah chatbot is a prototype of ALICE chatbot, using the same knowledge base (AIML) files of ALICE. Sarah was created to enable public chatting with it using the pandorabot host serving. The Loebner prize competition has been used to evaluate machine conversation chatbots. The Loebner Prize is a Turing test, which evaluates the ability of the machine to fool people that they are talking to human. In essence, judges are allowed a short chat (10 to 15 minutes) with each chatbot, and asked to rank them in terms of “naturalness”.

Quality Assessment for Text Simplification

28 May 2016

ABSTRACTS

Editors:

Sanja Štajner, Maja Popović, Horacio Saggion, Lucia Specia and Mark Fishel

Workshop Programme

28 May 2016

09:00 – 09:20 **Introduction** by Sanja Štajner

09:20 – 10:00 **Invited Talk** by Advaith Siddharthan

Session: General Track

10:00 – 10:30 Gustavo H. Paetzold and Lucia Specia
PLUMBErr: An Automatic Error Identification Framework for Lexical Simplification

10:30 – 11:00 Coffee break

11:00 – 11:30 Sandeep Mathias and Pushpak Bhattacharyya
How Hard Can it Be? The E-Score - A Scoring Metric to Assess the Complexity of Text

11:30 – 12:00 Sanja Štajner, Maja Popović and Hanna Béchara
Quality Estimation for Text Simplification

12:00 – 12:15 **Shared Task: Introduction** by Maja Popović

Session: Shared Task 1

12:15 - 12:45 Maja Popović and Sanja Štajner
Machine Translation Evaluation Metrics for Quality Assessment of Automatically Simplified Sentences

12:45 - 13:15 Sandeep Mathias and Pushpak Bhattacharyya
Using Machine Translation Evaluation Techniques to Evaluate Text Simplification Systems

13:15 - 14:30 Lunch break

Session: Shared Task 2

14:30 – 15:00	Gustavo H. Paetzold and Lucia Specia *SimpleNets: Evaluating Simplifiers with Resource-Light Neural Networks*
15:00 – 15:30	Sergiu Nisioi and Fabrice Nauze *An Ensemble Method for Quality Assessment of Text Simplification*
15:30 – 16:00	Elnaz Davoodi and Leila Kosseim *CLaC @ QATS: Quality Assessment for Text Simplification*
16:00 – 16:30	Coffee break
16:30 – 17:30	**Round Table**
17:30 – 17:45	**Closing**

Workshop Organizers

Sanja Štajner	University of Mannheim, Germany
Maja Popović	Humboldt University of Berlin, Germany
Horacio Saggion	Universitat Pompeu Fabra, Spain
Lucia Specia	University of Sheffield, UK
Mark Fishel	University of Tartu, Estonia

Workshop Programme Committee

Sandra Aluisio	University of São Paolo
Eleftherios Avramidis	DFKI Berlin
Susana Bautista	Federal University of Rio Grande do Sul
Stefan Bott	University of Stuttgart
Richard Evans	University of Wolverhampton
Mark Fishel	University of Tartu
Sujay Kumar Jahuar	Carnegie Mellon University
David Kauchak	Pomona College
Elena Lloret	Universidad de Alicante
Ruslan Mitkov	University of Wolverhampton
Gustavo Paetzold	University of Sheffield
Maja Popović	Humboldt University of Berlin
Miguel Rios	University of Leeds
Horacio Saggion	Universitat Pompeu Fabra
Carolina Scarton	University of Sheffield
Matthew Shardlow	University of Manchester
Advaith Siddharthan	University of Aberdeen
Lucia Specia	University of Sheffield, UK
Miloš Stanojević	University of Amsterdam
Sanja Štajner	University of Mannheim
Irina Temnikova	Qatar Computing Research Institute
Sowmya Vajjala	Iowa State University
Victoria Yaneva	University of Wolverhampton

Preface

In recent years, there has been an increasing interest in automatic text simplification (ATS) and text adaptation to various target populations. However, studies concerning evaluation of ATS systems are still very scarce and there are no methods proposed for directly comparing performances of different systems. This workshop addresses this problem and provides an opportunity to establish some metrics for automatic evaluation of ATS systems.

Given the close relatedness of the problem of automatic evaluation of ATS system to the well-studied problems of automatic evaluation and quality estimation in machine translation (MT), the workshop also features a shared task on automatic evaluation (quality assessment) of ATS systems.

We accepted three papers in the general track and five papers describing the systems which participated in the shared task. The papers describe a variety of interesting approaches to this task.

We wish to thank all people who helped in making this workshop a success. Our special thanks go to Advaith Siddharthan for accepting to give the invited presentation, to the members of the program committee who did an excellent job in reviewing the submitted papers and to the LREC organisers, as well as all authors and participants of the workshop.

Sanja Štajner, Maja Popović, Horacio Saggion, Lucia Specia and Mark Fishel

May 2016

Session: General Track

Saturday 28 May, 10:30 – 12:00
Chairperson: Advaith Siddharthan

PLUMBErr: An Automatic Error Identification Framework for Lexical Simplification

Gustavo H. Paetzold and Lucia Specia

Lexical Simplification is the task of replacing complex words with simpler alternatives. Using human evaluation to identify errors made by simplifiers throughout the simplification process can help to highlight their weaknesses, but is a costly process. To address this problem, we introduce PLUMBErr: an automatic alternative. Using PLUMBErr, we analyze over 40 systems, and find out the best combination to be the one between the winner of the Complex Word Identification task of SemEval 2016 and a modern simplifier. Comparing PLUMBErr to human judgments we find that, although reliable, PLUMBErr could benefit from resources annotated in a different way.

How Hard Can it Be? The E-Score - A Scoring Metric to Assess the Complexity of Text

Sandeep Mathias and Pushpak Bhattacharyya

In this paper, we present an evaluation metric, the E-Score, to calculate the complexity of text, that utilizes structural complexity of sentences and language modelling of simple and normal English to come up with a score that tells us how simple / complex the document is. We gather gold standard human data by having human participants take a comprehension test, in which they read articles from the English and Simple English Wikipedias. We use this data to evaluate our metric against a pair of popular existing metrics – the Flesch Reading Ease Score, and the Lexile Framework.

Quality Estimation for Text Simplification

Sanja Štajner, Maja Popović and Hanna Béchara

The quality of the output generated by automatic Text Simplification (TS) systems is traditionally assessed by human annotators. In spite of the fact that the automatisation of that process would enable faster and more consistent evaluation, there have been almost no studies addressing this problem. We propose several decision-making procedures for automatic classification of the simplified sentences into three classes (bad, OK, good) depending on their grammaticality, meaning preservation, and simplicity. We experiment with ten different classification algorithms and 12 different feature sets on three TS datasets obtained using different text simplification strategies, achieving the results significantly above the state of the art. Additionally, we propose to use an unique measure (Total2 or Total3) for classifying the quality of the automatically simplified sentences into two (discard or keep) or three (bad, OK, good) classes.

Session: Shared Task 1

Saturday 28 May, 12:15 – 13:15
Chairperson: Lucia Specia

Machine Translation Evaluation Metrics for Quality Assessment of Automatically Simplified Sentences

Maja Popović and Sanja Štajner

We investigate whether it is possible to automatically evaluate the output of automatic text simplification (ATS) systems by using automatic metrics designed for evaluation of machine translation (MT) outputs. In the first step, we select a set of the most promising metrics based on the Pearson's correlation coefficients between those metrics and human scores for the overall quality of automatically simplified sentences. Next, we build eight classifiers on the training dataset using the subset of 13 most promising metrics as features, and apply two best classifiers on the test set. Additionally, we apply an attribute selection algorithm to further select best subset of features for our classification experiments. Finally, we report on the success of our systems in the shared task and report on confusion matrices which can help to gain better insights into the most challenging problems of this task.

Using Machine Translation Evaluation Techniques to Evaluate Text Simplification Systems

Sandeep Mathias and Pushpak Bhattacharyya

In this paper, we discuss our approaches to find out ways to evaluate automated text simplification systems, based on the grammaticality and simplicity of their output, as well as the meaning preserved from the input, and the overall quality of simplification of the system. In this paper, we discuss existing techniques currently used in the area of machine translation, as well as a novel technique for text complexity analysis, to assess the quality of the text simplification system.

Session: Shared Task 2

Saturday 28 May, 14:30 – 16:00
Chairperson: Sanja Štajner

SimpleNets: Evaluating Simplifiers with Resource-Light Neural Networks

Gustavo H. Paetzold and Lucia Specia

We present our contribution to the shared task on Quality Assessment for Text Simplification at QATS 2016: the SimpleNets systems. We introduce a resource-light Multi-Layer Perceptron classifier, as well as a deep Recurrent Neural Network that predicts the quality of a simplification by assessing the quality of its n-grams. Our Recurrent Neural Networks have achieved the state-of-the-art solution for the task, outperforming all other system in terms of overall simplification Accuracy.

An Ensemble Method for Quality Assessment of Text Simplification

Sergiu Nisioi and Fabrice Nauze

In this paper we describe the Oracle Service Cloud Machine Learning (OSVCML) systems used for the Quality Assessment of Text Simplification, 2016 (QATS) shared task. We construct an ensemble method using particle swarm optimization and different scoring methods (SVM, string kernels, logistic regression, boosting trees, BLEU). The purpose is to capture relevant combinations of classifier and features for each different aspects of text simplification: simplicity, grammaticality, meaning preservation, and overall scores. In addition, we compare our approach with a deep neural network architecture and show that the generated models are stronger when combined together.

CLaC @ QATS: Quality Assessment for Text Simplification

Elnaz Davoodi and Leila Kosseim

This paper describes our approach to the 2016 QATS quality assessment shared task. We trained three independent Random Forest classifiers in order to assess the quality of the simplified texts in terms of grammaticality, meaning preservation and simplicity. We used the language model of Google-Ngram as feature to predict the grammaticality. Meaning preservation is predicted using two complementary approaches based on word embedding and WordNet synonyms. A wider range of features including TF-IDF, sentence length and frequency of cue phrases are used to evaluate the simplicity aspect. Overall, the accuracy of the system ranges from 33.33% for the overall aspect to 58.73% for grammaticality.