COLING 2014

The 25th International Conference
on Computational Linguistics

Proceedings of the
Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014)

August 24, 2014
Dublin, Ireland

©2014 Copyright of each paper stays with the respective authors. The works in the Proceedings are licensed under a Creative Commons Attribution 4.0 International Licence. License details: http://creativecommons.org/licenses/by/4.0

ISBN 978-1-873769-44-7

Introduction

The first instance of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014) took place on August 24th in Dublin, in conjunction with COLING 2014. It was co-sponsored by ASIALEX and endorsed by SIGLEX.

The workshop aimed to bring together members of the language-resource (LR) landscape, focusing on complex linguistic knowledge that requires linguistic expertise, e.g. on dictionaries, ontologies and grammars. Such manually-built resources are key to the development of natural language processing (NLP) tools and applications. We intended to strengthen the cohesion of the scientific 'production chain' spanning from the construction of LRs to their exploitation in hybrid or symbolic NLP. It is necessary to increase mutual awareness between researchers along this production chain, regarding their activities, skills and needs, in view of improving the building processes of the resources, their validation and their exploitation.

Many linguists are comfortable with descriptive tasks such as checking lexical entries for a given feature, even if each entry requires analysing or pondering. On the other hand, computer scientists are familiar with formalization and, usually, with notions such as falsifiability or reproducibility, which are fundamental to sciences. Combining all these skills is likely to stimulate innovation. The workshop offered an opportunity of interaction which is required to overcome the compartmentalization between humanities and sciences, and to intensify co-operation between the two ends of the chain. Researchers were encouraged to exchange about how they manage to face several challenges:

- the context of this production chain requires that they not be content with understanding phenomena, but also achieve actual production of formalized results;
- resulting resources should reach a reasonable level of verifiability, e.g. by finding formal or syntactic bases as a support to semantic description;
- methods which are able to cover the most diverse languages are to be preferred;
- the format of manual construction of complex LRs must be highly readable, so that errors can be easily detected and corrected;
- conceptual models are not easy to assign to large amounts of language data; due to idiosyncratic behaviour of lexical entries, it is often required to manually examine them individually as regards syntax or semantics;
- many multiword expressions, including support-verb constructions, are somewhere halfway between compositional and non-compositional constructs;
- actual implementation of NLP systems and real-world applications may provide feedback on complex lexical and grammatical LRs used in them, but experimentation is required to accurately relate features of the LRs with features of results obtained in NLP.

We received 31 submissions and accepted 19: an acceptance rate of 61%. We scheduled 10 papers for oral presentation and 9 as posters. The workshop closed with a general discussion.

We would like to thank the members of the Program Committee for their timely reviews. We would also like to thank the authors for their valuable contributions.

Jorge Baptista, Pushpak Bhattacharyya, Christiane Fellbaum, Mikel Forcada, Chu-Ren Huang, Svetla Koeva, Cvetana Krstev, Éric Laporte
Co-Organizers

Organizers:

Jorge Baptista, University of Algarve, Portugal
Pushpak Bhattacharyya, Indian Institute of Technology Bombay, India
Christiane Fellbaum, Princeton University, USA
Mikel Forcada, Universitat d'Alacant, Spain
Chu-Ren Huang, The Hong Kong Polytechnic University, Hong-Kong
Svetla Koeva, Bulgarian Academy of Sciences, Bulgaria
Cvetana Krstev, University of Belgrade, Serbia
Éric Laporte, Université Paris-Est Marne-la-Vallée, France

Program Committee:

Wirote Arunmanakun, Chulalongkorn University, Thailand
Jorge Baptista, University of Algarve, Portugal
Núria Bel, Universitat Pompeu Fabra, Spain
Pushpak Bhattacharyya, Indian Institute of Technology Bombay, India
Dunstan Brown, University of York, UK
Rebecca Dridan, University of Oslo, Norway
Christiane Fellbaum, Princeton University, US
Mikel Forcada, University of Alicante, Spain
Chu-Ren Huang, Polytechnic University, Hong-Kong
Svetla Koeva, Bulgarian Academy of Sciences, Bulgaria
Cvetana Krstev, University of Belgrade, Serbia
Éric Laporte, Université Paris-Est Marne-la-Vallée, France
Nuno Mamede, IST-UL, Portugal
Ruli Manurung, University of Indonesia, Indonesia
Denis Maurel, Université de Tours, France
Nurit Melnik, Open University, Israel
Adam Meyers, New York University, US
Jee-sun Nam, Hankuk University of Foreign Studies, Korea
Maria das Graças Volpe Nunes, Universidade de São Paulo, Brazil
Kemal Oflazer, Carnegie-Mellon University, Qatar
Thiago Pardo, Universidade de São Paulo, Brazil
Adam Pease, Articulate Software and the Hong Kong Polytechnic University, US & Hong Kong
Miriam Petruck, International Computer Science Institute, Berkeley, US
Adam Przepiórkowski, Polish Academy of Sciences, Poland
Laurent Romary, Humboldt University of Berlin, Germany
Rachel E. Roxas, De LaSalle University, the Philippines
Agata Savary, Université de Tours, France
Carlos Subirats, Universidad Autonoma de Barcelona, Spain
Yukio Tono, Tokyo University of Foreign Studies, Japan
Francis M. Tyers, Noregs Arktiske Universitet, Tromsø, Norway
Aline Villavicencio, Universidade Federal do Rio Gande do Sul, Brazil

Revision of the proceedings:

Takuya Nakamura, LIGM, CNRS, France

Table of Contents

Paraphrasing of Italian Support Verb Constructions based on Lexical and Grammatical Resources
Konstantinos Chatzitheodorou 1

Using language technology resources and tools to construct Swedish FrameNet
Dana Dannells, Karin Friberg Heppin and Anna Ehrlemark 8

Harmonizing Lexical Data for their Linking to Knowledge Objects in the Linked Data Framework
Thierry Declerck 18

Terminology and Knowledge Representation. Italian Linguistic Resources for the Archaeological Domain
Maria Pia di Buono, Mario Monteleone and Annibale Elia 24

SentiMerge: Combining Sentiment Lexicons in a Bayesian Framework
Guy Emerson and Thierry Declerck 30

Linguistically motivated Language Resources for Sentiment Analysis
Voula Giouli and Aggeliki Fotopoulou 39

Using Morphosemantic Information in Construction of a Pilot Lexical Semantic Resource for Turkish
Gözde Gül İşgüder and Eşref Adalı 46

Comparing Czech and English AMRs
Jan Hajic, Ondrej Bojar and Zdenka Uresova 55

Acquisition and enrichment of morphological and morphosemantic knowledge from the French Wiktionary
Nabil Hathout, Franck Sajous and Basilio Calderone 65

Annotation and Classification of Light Verbs and Light Verb Variations in Mandarin Chinese
Jingxia Lin, Hongzhi Xu, Menghan JIANG and Chu-Ren Huang 75

Extended phraseological information in a valence dictionary for NLP applications
Adam Przepiórkowski, Elżbieta Hajnicz, Agnieszka Patejuk and Marcin Woliński 84

The fuzzy boundaries of operator verb and support verb constructions with dar "give" and ter "have" in Brazilian Portuguese
Amanda Rassi, Cristina Santos-Turati, Jorge Baptista, Nuno Mamede and Oto Vale 93

Collaboratively Constructed Linguistic Resources for Language Variants and their Exploitation in NLP Application – the case of Tunisian Arabic and the Social Media
Fatiha Sadat, Fatma Mallek, Mohamed Boudabous, Rahma Sellami and Atefeh Farzindar 103

A Database of Paradigmatic Semantic Relation Pairs for German Nouns, Verbs, and Adjectives
Silke Scheible and Sabine Schulte im Walde 112

Improving the Precision of Synset Links Between Cornetto and Princeton WordNet
Leen Sevens, Vincent Vandeghinste and Frank Van Eynde 121

Light verb constructions with 'do' and 'be' in Hindi: A TAG analysis
Ashwini Vaidya, Owen Rambow and Martha Palmer 128

The Lexicon-Grammar of Italian Idioms
Simonetta Vietri 138

Building a Semantic Transparency Dataset of Chinese Nominal Compounds: A Practice of Crowdsourcing Methodology
Shichang Wang, Chu-Ren Huang, Yao Yao and Angel Chan 148

Annotate and Identify Modalities, Speech Acts and Finer-Grained Event Types in Chinese Text
Hongzhi Xu and Chu-Ren Huang 158