EACL 2014

14th Conference of the European Chapter of the Association for Computational Linguistics

Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)

April 26, 2014
Gothenburg, Sweden

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-937284-85-5

Preface

The LaTeCH workshop series, which started in 2007, was initially motivated by the growing interest in language technology research and applications to the cultural heritage domain. The scope quickly broadened to also include the humanities and the social sciences. LaTeCH is currently the annual venue of the ACL Special Interest Group on Language Technologies for the Socio-Economic Sciences and Humanities (SIGHUM).

In the current, eighth edition of the LaTeCH workshop, we have received a record number of submissions, a subset of which has been selected based on a thorough peer-review process. The submissions were substantial not only in terms of quantity, but also in terms of quality and variety, underlining the interest of NLP and CL researchers in this exciting and expanding research area.

For this edition of LaTeCH, we attempted to focus on *Linked data in the Humanities*, an issue also addressed by our invited speaker, Gerhard Heyer in his talk about the Canonical Text Services protocol implementations in the digital humanities. Linked data has fairly recently regained a particular research interest in our field, as also indicated by the respective contributions to LaTeCH-2014. Apart for the recurring themes of linguistic variability in historical text, OCR error correction and annotation tools and resource development, we were delighted in this edition of our workshop to receive contributions about applications in social sciences and resource development for non-European languages and cultural heritage, such as the work on the Tagalog Linguistic Inquiry Dictionary, a dictionary for disaster terms in the Tagalog language of Philippines, and the work on the development of a wayang ontology, an ontology about the Indonesian shadow puppet mythology. The acceptance rate for LaTeCH-2014 was 68%.

We would like to thank all authors for the hard work that went into their submissions. We are also grateful to the members of the programme committee for their thorough reviews, and to the EACL 2014 organisers, especially the Workshop Co-chairs, Anja Belz and Reut Tsarfaty for their help with administrative matters.

Kalliopi Zervanou and Cristina Vertan

Organizers:

Kalliopi Zervanou (Co-Chair), Radboud University Nijmegen (The Netherlands)
Cristina Vertan (Co-Chair), University of Hamburg (Germany)
Antal van den Bosch, Radboud University Nijmegen (The Netherlands)
Caroline Sporleder, Trier University (Germany)

Program Committee:

Laura Alonso Alemany, Universidad Nacional de Cordoba (Argentina)
Ion Androutsopoulos, Athens University of Economics and Business (Greece)
Andrei Beliankou, Trier University (Germany)
Kristín Bjarnadóttir, Àrni Magnússon Institute for Icelandic Studies (Iceland)
Toine Bogers, Aalborg University, Copenhagen (Denmark)
Paul Buitelaar, DERI Galway (Ireland)
Mariona Coll Ardanuy, Trier University (Germany)
Thierry Declerck, DFKI (Germany)
Stefanie Dipper, Ruhr-Universität Bochum (Germany)
Milena Dobreva, University of Malta (Malta)
Mick O'Donnell, Universidad Autonoma de Madrid (Spain)
Ben Hachey, Macquarie University (Australia)
Iris Hendrickx, Radboud University Nijmegen (The Netherlands)
Elias Iosif, Technical University of Crete (Greece)
Jaap Kamps, University of Amsterdam (The Netherlands)
Vangelis Karkaletsis, NCSR Demokritos (Greece)
Mike Kestemont, University of Antwerp & Research Foundation Flanders (Belgium)
Dimitrios Kokkinakis, University of Gothenburg (Sweden)
Stasinos Konstantopoulos, NCSR Demokritos (Greece)
Piroska Lendvai, cliqz (Germany)
Barbara McGillivray, Oxford University Press
Joakim Nivre, Uppsala University (Sweden)
Nelleke Oostdijk, Radboud University Nijmegen (The Netherlands)
Csaba Oravecz Research Institute for Linguistics (HASRIL) (Hungary)
Petya Osenova, Bulgarian Academy of Sciences (Bulgaria)
Katerina Pastra, Cognitive Systems Research Institute (CSRI) (Greece)
Michael Piotrowski, Leibniz Institute of European History in Mainz (Germany)
Georg Rehm, DFKI (Germany)
Martin Reynaert, Tilburg University (The Netherlands)
Eric Sanders, Radboud University Nijmegen (The Netherlands)
Eszter Simon, Research Institute for Linguistics (HASRIL) (Hungary)
Herman Stehouwer, Max Planck for Plasmaphysics (Germany)
Mark Stevenson, University of Sheffield (UK)
Mariët Theune, University of Twente (The Netherlands)
Suzan Verberne, Radboud University Nijmegen (The Netherlands)
Manolis Wallace, University of Peloponnese (Greece)
Menno van Zaanen, Tilburg University (The Netherlands)
Svitlana Zinger, TU Eindhoven (The Netherlands)

Table of Contents

Workshop Program

Saturday, April 26, 2014

8:45–8:50 Welcome

8:50–9:30 Invited Talk by Gerhard Heyer:
A New Implementation for Canonical Text Services
Jochen Tiepmar, Christoph Teichmann, Gerhard Heyer, Monica Berti and Gregory Crane

Session I: Linked data in the Humanities

9:30–9:45 *How to semantically relate dialectal Dictionaries in the Linked Data Framework*
Thierry Declerck and Eveline Wandl-Vogt

9:45–10:05 *Bootstrapping a historical commodities lexicon with SKOS and DBpedia*
Ewan Klein, Beatrice Alex and Jim Clifford

10:05–10:25 *New Technologies for Old Germanic. Resources and Research on Parallel Bibles in Older Continental Western Germanic*
Christian Chiarcos, Maria Sukhareva, Roland Mittmann, Timothy Price, Gaye Detmold and Jan Chobotsky

10:25–11:00 Coffee break

Session II: Spelling normalisation & sense disambiguation

11:00–11:20 *A Multilingual Evaluation of Three Spelling Normalisation Methods for Historical Text*
Eva Pettersson, Beáta Megyesi and Joakim Nivre

11:20–11:40 *Enhancing the possibilities of corpus-based investigations: Word sense disambiguation on query results of large text corpora*
Christian Poelitz and Thomas Bartz

11:40–12:00 *A Hybrid Disambiguation Measure for Inaccurate Cultural Heritage Data*
Julia Efremova, Bijan Ranjbar-Sahraei and Toon Calders

12:00–12:15 *Automated Error Detection in Digitized Cultural Heritage Documents*
Kata Gábor and Benoît Sagot

12:15–13:45 Lunch break

Saturday, April 26, 2014 (continued)

Session III: Social Science applications

13:45–14:05 *Mining the Twentieth Century's History from the Time Magazine Corpus*
Mike Kestemont, Folgert Karsdorp and Marten Düring

14:05–14:25 *Social and Semantic Diversity:*
Socio-semantic Representation of a Scientific Corpus
Thierry Poibeau, Elisa Omodei, Jean-Philippe Cointet and Yufan Guo

Poster Booster Session

14:25–14:35 *A Tool for a High-Carat Gold-Standard Word Alignment*
Drayton Benner

14:35–14:45 *CorA: A web-based annotation tool for historical and other non-standard language data*
Marcel Bollmann, Florian Petran, Stefanie Dipper and Julia Krasselt

14:45–14:55 *Developing a Tagalog Linguistic Inquiry and Word Count (LIWC) 'Disaster' Dictionary for Understanding Mixed Language Social Media: A Work-in-Progress Paper*
Amanda Andrei, Alison Dingwall, Theresa Dillon and Jennifer Mathieu

14:55–15:05 *Text Analysis of Aberdeen Burgh Records 1530-1531*
Adam Wyner, Jackson Armstrong, Andrew Mackillop and Philip Astley

15:05–15:15 *From Syntax to Semantics. First Steps Towards Tectogrammatical Annotation of Latin*
Marco Passarotti

15:15–15:25 *On the syllabic structures of Aromanian*
Sergiu Nisioi

15:25–16:00 Coffee break & Poster Session

Session IV: Knowledge resources acquisition

16:00–16:20 *A Gazetteer and Georeferencing for Historical English Documents*
Claire Grover and Richard Tobin

16:20–16:40 *Automatic Wayang Ontology Construction using Relation Extraction from Free Text*
Hadaiq Sanabila and Ruli Manurung

16:40–17:30 SIGHUM annual business meeting