EACL 2012

Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2012)

24 April 2012
Avignon, France

ISBN 978-1-937284-19-0

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

Preface

The LaTeCH (*Language Technology for Cultural Heritage, Social Sciences, and Humanities*) annual workshop series aims to provide a forum for researchers working on aspects of natural language and information technology applications that pertain to data from the humanities, social sciences, and cultural heritage. 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.

Advances in information technology and web access of the past decade have triggered a multitude of digitisation efforts by museums, archives, libraries and other cultural heritage institutions. Similar developments in the humanities and social sciences have resulted in large amounts of research data becoming available in electronic format, either as digitised or as born-digital data. The natural follow-up step to digitisation is the intelligent processing of this data. To this end, the humanities, social sciences, and cultural heritage domains draw an increasing interest from researchers in NLP aiming at developing methods for semantic enrichment and information discovery and access. Language technology has been conventionally focused on certain domains, such as newswire. These fairly novel domains of cultural heritage, social sciences, and humanities entail new challenges to NLP research, such as noisy text (e.g., due to OCR problems), non-standard, or archaic language varieties (e.g., historic language, dialects, mixed use of languages, ellipsis, transcription errors), literary or figurative writing style and lack of knowledge resources, such as dictionaries. Furthermore, often neither annotated domain data is available, nor the required funds to manually create it, thus forcing researchers to investigate (semi-) automatic resource development and domain adaptation approaches involving the least possible manual effort.

In the current sixth edition of the LaTeCH workshop we have again 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 increasing interest of NLP and CL researchers in in this exciting and expanding research area. A central issue for a substantial part of the contributions to this LaTeCH workshop is the development of corpora, lexical resources and annotation tools for historical language varieties. Some contributions focus on two other recurrent issues, namely the description and classification of cultural heritage objects and the problems related to noisy and handwritten text. Most importantly, the research presented in this edition of the LaTeCH workshop showcases the breadth of interest in applications of language technologies to a variety of social sciences and e-humanities domains, ranging from history to ethnography and from philosophy to literature and historical linguistics.

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 2012 organisers, especially the Workshop Co-chairs, Kristiina Jokinen and Alessandro Moschitti for their help with administrative matters.

Kalliopi Zervanou and Antal van den Bosch

Organizers:

Kalliopi Zervanou (Co-chair), Tilburg University (The Netherlands)
Antal van den Bosch (Co-chair), Radboud University Nijmegen (The Netherlands)
Caroline Sporleder, Saarland University (Germany)
Piroska Lendvai, Research Institute for Linguistics (Hungary)

Program Committee:

Ion Androutsopoulos, Athens University of Economics and Business (Greece)
David Bamman, Carnegie Mellon University (USA)
Toine Bogers, Royal School of Library & Information Science, Copenhagen (Denmark)
Paul Buitelaar, DERI Galway (Ireland)
Kate Byrne, University of Edinburgh (Scotland)
Milena Dobreva, HATII, University of Glasgow (Scotland)
Mick O'Donnell, Universidad Autonoma de Madrid (Spain)
Claire Grover, University of Edinburgh (Scotland)
Ben Hachey, Macquarie University (Australia)
Jaap Kamps, University of Amsterdam (The Netherlands)
Vangelis Karkaletsis, NCSR Demokritos (Greece)
Stasinos Konstantopoulos, NCSR Demokritos (Greece)
Barbara McGillivray, Oxford University Press
Joakim Nivre, Uppsala University (Sweden)
Petya Osenova, Bulgarian Academy of Sciences (Bulgaria)
Katerina Pastra, CSRI (Greece)
Michael Piotrowski, University of Zurich (Switzerland)
Georg Rehm, DFKI (Germany)
Martin Reynaert, Tilburg University (The Netherlands)
Herman Stehouwer, Max Planck Institute for Psycholinguistics (The Netherlands)
Cristina Vertan, University of Hamburg (Germany)
Piek Vossen, Vrije Universiteit Amsterdam (The Netherlands)
Peter Wittenburg, Max Planck Institute for Psycholinguistics (The Netherlands)
Menno van Zaanen, Tilburg University (The Netherlands)
Svitlana Zinger, TU Eindhoven (The Netherlands)

Table of Contents

Conference Program

Tuesday April 24, 2012

9:00–9:05 *Welcome*

Poster Boaster Session: Tools & Resources

9:05–9:10 *Lexicon Construction and Corpus Annotation of Historical Language with the CoBaLT Editor*
Tom Kenter, Tomaž Erjavec, Maja Žorga Dulmin and Darja Fišer

9:10–9:15 *A High Speed Transcription Interface for Annotating Primary Linguistic Data*
Mark Dingemanse, Jeremy Hammond, Herman Stehouwer, Aarthy Somasundaram and Sebastian Drude

9:15–9:20 *BAD: An Assistant Tool for Making Verses in Basque*
Manex Agirrezabal, Iñaki Alegria, Bertol Arrieta and Mans Hulden

9:20–9:25 *Toward Language Independent Methodology for Generating Artwork Descriptions – Exploring FrameNet Information*
Dana Dannélls and Lars Borin

9:25–9:30 *Harvesting Indices to Grow a Controlled Vocabulary: Towards Improved Access to Historical Legal Texts*
Michael Piotrowski and Cathrin Senn

9:30–9:35 *Ontology-Based Incremental Annotation of Characters in Folktales*
Thierry Declerck, Nikolina Koleva and Hans-Ulrich Krieger

9:35–10:30 Poster session & Coffie break

Oral Session 1: Applications in Humanities & Social Sciences

10:30–11:00 *Advanced Visual Analytics Methods for Literature Analysis*
Daniela Oelke, Dimitrios Kokkinakis and Mats Malm

11:00–11:30 *Distributional Techniques for Philosophical Enquiry*
Aurélie Herbelot, Eva von Redecker and Johanna Müller

11:30–12:00 *Linguistically-Adapted Structural Query Annotation for Digital Libraries in the Social Sciences*
Caroline Brun, Vassilina Nikoulina and Nikolaos Lagos

12:00–12:30 *Parsing the Past – Identification of Verb Constructions in Historical Text*
Eva Pettersson, Beáta Megyesi and Joakim Nivre

Tuesday April 24, 2012 (continued)

12:30–14:00 Lunch break

Oral Session 2: Cultural Heritage Objects

14:00–14:30 *A Classical Chinese Corpus with Nested Part-of-Speech Tags*
John Lee

14:30–15:00 *Computing Similarity between Cultural Heritage Items using Multimodal Features*
Nikolaos Aletras and Mark Stevenson

15:00–15:20 *Enabling the Discovery of Digital Cultural Heritage Objects through Wikipedia*
Mark Michael Hall, Oier Lopez de Lacalle, Aitor Soroa Etxabe, Paul Clough and Eneko Agirre

15:20–15:40 *Adapting Wikification to Cultural Heritage*
Samuel Fernando and Mark Stevenson

15:40–16:10 Coffee break

Oral Session 3: Historical & Handwritten Documents

16:10–16:30 *Natural Language Inspired Approach for Handwritten Text Line Detection in Legacy Documents*
Vicente Bosch, Alejandro Héctor Toselli and Enrique Vidal

16:30–16:50 *Language Classification and Segmentation of Noisy Documents in Hebrew Scripts*
Alex Zhicharevich and Nachum Dershowitz

Discussion Session

16:50–17:30 *Towards a Special Interest Group in Language Technology for the Humanities*
Kalliopi Zervanou, Caroline Sporleder and Antal van den Bosch