ACL HLT 2011

Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities LaTeCH

Proceedings of the Workshop

24 June, 2011
Portland, Oregon, USA

Production and Manufacturing by
Omnipress, Inc.
2600 Anderson Street
Madison, WI 53704 USA

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-13 9781937284046

Preface

The LaTeCH (*Language Technology for Cultural Heritage, Social Sciences, and Humanities*) annual workshop series aims to provide a forum for researchers who are working on aspects of natural language and information technology applications that pertain to data from the humanities, social sciences, and cultural heritage. The LaTeCH workshops were initially motivated by the growing interest in language technology research and applications for the cultural heritage domain. The scope has soon nevertheless broadened to also include the humanities and the social sciences.

Current developments in web and information access have triggered a series of digitisation efforts by museums, archives, libraries and other cultural heritage institutions. Similar developments in humanities and social sciences have resulted in large amounts of data becoming available in electronic format, either as digitised, or as born-digital data. The natural next 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 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. A central issue for the majority of contributions to this LaTeCH workshop has been the problem of linguistic processing for historical language varieties (e.g., Spanish, Czech, German, Slovene and Swedish) and the respective resource development and tool adaptation. In terms of applications, the contributions attempt to provide language technology solutions for cultural heritage and humanities researchers ranging from historians and architecture historians to linguists, cultural heritage curators, ethnologists and literary critics. The text types targeted for analysis range from full-text to semi-structured text, while the domains addressed range from the analysis of historical text and encrypted medieval manuscripts, to novels and fairy tales and modern academic journals, online blogs and fora. The variety of topics and the increased number of submissions illustrate the growing interest in this exciting and expanding research area.

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 ACL-HLT 2011 organisers, especially the Workshop Co-chairs, Hal Daumé III and John Carroll for their help with administrative matters.

Kalliopi Zervanou & Piroska Lendvai

Organizers:

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

Program Committee:

Ion Androutsopoulos, Athens University of Economics and Business (Greece)
Tim Baldwin, University of Melbourne (Australia)
David Bamman, Tufts 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)
Julio Gonzalo, Universidad Nacional de Educacion a Distancia (Spain)
Claire Grover, University of Edinburgh (Scotland)
Ben Hachey, Macquarie University (Australia)
Eduard Hovy, USC Information Sciences Institute (USA)
Jaap Kamps, University of Amsterdam (The Netherlands)
Vangelis Karkaletsis, NCSR Demokritos (Greece)
Stasinos Konstantopoulos, NCSR Demokritos (Greece)
Ioannis Korkontzelos, National Centre for Text Mining – NaCTeM (UK)
Véronique Malaisé, Elsevier, Content Enrichment Center
Barbara McGillivray, Oxford University Press
John Nerbonne, Rijksuniversiteit Groningen (The Netherlands)
Katerina Pastra, CSRI (Greece)
Michael Piotrowski, University of Zurich (Switzerland)
Georg Rehm, DFKI (Germany)
Martin Reynaert, University of Tilburg (The Netherlands)
Svitlana Zinger, TU Eindhoven (The Netherlands)

Table of Contents

Conference Program

Friday June 24, 2011

9:00–9:10	Welcome
9:10–9:40	*Extending the tool, or how to annotate historical language varieties* Cristina Sánchez-Marco, Gemma Boleda and Lluís Padró
9:40–10:10	*A low-budget tagger for Old Czech* Jirka Hana, Anna Feldman and Katsiaryna Aharodnik
10:10–10:30	*Evaluating an ‘off-the-shelf’ POS-tagger on Early Modern German text* Silke Scheible, Richard J. Whitt, Martin Durrell and Paul Bennett
10:30–11:00	Coffee break
11:00–11:10	*e-Research for Linguists* Dorothee Beermann and Pavel Mihaylov
11:10–11:15	*Automatic linguistic annotation of historical language: ToTrTaLe and XIX century Slovene* Tomaž Erjavec
11:15–11:20	*Historical Event Extraction from Text* Agata Katarzyna Cybulska and Piek Vossen
11:20–11:30	*Enrichment and Structuring of Archival Description Metadata* Kalliopi Zervanou, Ioannis Korkontzelos, Antal van den Bosch and Sophia Ananiadou
11:30–11:40	*Structure-Preserving Pipelines for Digital Libraries* Massimo Poesio, Eduard Barbu, Egon Stemle and Christian Girardi
11:40–11:45	*The ARC Project: Creating logical models of Gothic cathedrals using natural language processing* Charles Hollingsworth, Stefaan Van Liefferinge, Rebecca A. Smith, Michael A. Covington and Walter D. Potter
11:45–11:55	*Crowdsourcing syntactic relatedness judgements for opinion mining in the study of information technology adoption* Asad Sayeed, Bryan Rusk, Martin Petrov, Hieu Nguyen, Timothy Meyer and Amy Weinberg
12:00–13:00	Poster Session
13:00–14:00	Lunch break

Friday June 24, 2011 (continued)

14:00–14:30	*What We Know About The Voynich Manuscript* Sravana Reddy and Kevin Knight
14:30–15:00	*Automatic Verb Extraction from Historical Swedish Texts* Eva Pettersson and Joakim Nivre
15:00–15:30	*Topic Modeling on Historical Newspapers* Tze-I Yang, Andrew Torget and Rada Mihalcea
15:30–16:00	Coffee break
16:00–16:30	*From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales* Saif Mohammad
16:30–17:00	*Author Age Prediction from Text using Linear Regression* Dong Nguyen, Noah A. Smith and Carolyn P. Rosé
17:00–17:30	*A Study of Academic Collaborations in Computational Linguistics using a Latent Mixture of Authors Model* Nikhil Johri, Daniel Ramage, Daniel McFarland and Daniel Jurafsky