ACL 2007

PRAGUE

ACL 2007

Proceedings of the Workshop on Language Technology for Cultural Heritage Data

June 28, 2007
Prague, Czech Republic

The Association for Computational Linguistics

Production and Manufacturing by
Omnipress
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`

Preface

Museums, archives, and libraries around the world maintain large collections of cultural heritage objects, such as archaeological artefacts, sound recordings, historical manuscripts, or preserved animal specimens. Large scale digitisation projects are currently underway to make these collections more accessible. The natural next step after digitisation is the development of powerful tools to search, link, enrich, and mine the digitised data. Language technology has an important role to play in this endeavour, even for collections which are primarily non-textual, since text is the pervasive medium used for metadata. At the same time, the cultural heritage domain poses special challenges for the NLP community, including the use of historical or non-standard language and orthography, the presence of OCR or transcription errors in the input data, and the necessity to deal with data from various media and languages. The cultural heritage domain is therefore also a challenging and interesting testbed for the robustness of existing language technology.

The ACL 2007 workshop on *Language Technology for Cultural Heritage Data* is to be seen in the context of a growing interest in the development of IT solutions for the cultural heritage domain, as witnessed by numerous national and international research initiatives, such as CATCH (Continuous Access to Cultural Heritage), DigiCULT (Digital Culture), MALACH (Multilingual Access to Large Spoken Archives), and MultiMatch (Multilingual/Multimedia Access To Cultural Heritage).

We solicited papers describing new and original work on all aspects of language technology for the cultural heritage domain. Out of the 22 submissions received, 11 were selected for inclusion in the workshop programme following a peer-review process. The list of papers reflects the current breadth of this exciting and expanding area, with topics covering improved access to cultural heritage data (combining digital libraries with treebanks, mono- and cross-lingual information retrieval, dealing with controlled vocabularies), methods for aligning hand-written documents with their transcripts, named entity recognition for historical texts, knowledge discovery in databases, and museum visitor path prediction. An invited talk by Douglas W. Oard on the MALACH project completes the workshop programme.

We would like to thank all authors who submitted papers for the hard work that went into their submissions. We are also extremely grateful to the members of the programme committee for their thorough reviews, and to the ACL 2007 organisers, especially the ACL 2007 Workshop Chair Simone Teufel, for their help with administrative matters. Special thanks to our invited speaker Doug Oard and to the MultiMatch project for their generous sponsorship of the workshop.

Antal van den Bosch
Claire Grover
Caroline Sporleder

Organizers

Chairs:

Caroline Sporleder, Saarland University
Antal van den Bosch, University of Tilburg
Claire Grover, University of Edinburgh

Program Committee:

Ion Androutsopoulos, Athens University of Economics and Business
Antal van den Bosch, Tilburg University
Kate Byrne, University of Edinburgh
Robert Dale, Macquarie University
Vania Dimitrova, University of Leeds
Mick O'Donnell, Universidad Autonoma de Madrid
Bassilis Gatos, NCSR Demokritos
Julio Gonzalo, Universidad Nacional de Educacion a Distancia
Claire Grover, University of Edinburgh
Jiyin He, University of Amsterdam
Marti Hearst, University of California Berkeley
Djoerd Hiemstra, University of Twente
Nancy Ide, Vassar College
Neil Ireson, University of Sheffield
Christer Johansson, University of Bergen
Franciska de Jong, University of Twente
Jaap Kamps, University of Amsterdam
Vangelis Karkaletsis, NCSR Demokritos
Piroska Lendvai, Tilburg University
Ruli Manurung, University of Indonesia
Maria Milosavljevic, University of Edinburgh
Marie-Francine Moens, Katholieke Universiteit Leuven
John Nerbonne, Rijksuniversiteit Groningen
Douglas Oard, University of Maryland
Hans Paijmans, Maastricht University
Martin Reynaert, Tilburg University
Maarten de Rijke, University of Amsterdam
Mark Sanderson, University of Sheffield
Caroline Sporleder, Saarland University
Efstathios Stamatatos, University of the Aegean
Erik Tjong Kim Sang, University of Amsterdam
Arjen de Vries, CWI, Amsterdam

Invited Speaker:

Douglas W. Oard, University of Maryland

Table of Contents

Conference Program

Thursday, June 28, 2007

9:00–9:05 Welcome

9:05–9:30 *Naming the Past: Named Entity and Animacy Recognition in 19th Century Swedish Literature*
Lars Borin, Dimitrios Kokkinakis and Leif-Jöran Olsson

9:30–9:55 *Viterbi Based Alignment between Text Images and their Transcripts*
Alejandro H. Toselli, Verónica Romero and Enrique Vidal

9:55–10:20 *Retrieving Lost Information from Textual Databases: Rediscovering Expeditions from an Animal Specimen Database*
Marieke van Erp

10:20–10:45 *Concept Disambiguation for Improved Subject Access Using Multiple Knowledge Sources*
Tandeep Sidhu, Judith Klavans and Jimmy Lin

10:45–11:15 Coffee Break and Poster Session

The Latin Dependency Treebank in a Cultural Heritage Digital Library
David Bamman and Gregory Crane

Cultural Heritage Digital Resources: From Extraction to Querying
Michel Généreux

Dynamic Path Prediction and Recommendation in a Museum Environment
Karl Grieser, Timothy Baldwin and Steven Bird

Anchoring Dutch Cultural Heritage Thesauri to WordNet: Two Case Studies
Véronique Malaisé, Antoine Isaac, Luit Gazendam and Hennie Brugman

Cross Lingual and Semantic Retrieval for Cultural Heritage Appreciation
Idan Szpektor, Ido Dagan, Alon Lavie, Danny Shacham and Shuly Wintner

11:15–11:40 *Deriving a Domain Specific Test Collection from a Query Log*
Avi Arampatzis, Jaap Kamps, Marijn Koolen and Nir Nussbaum

Thursday, June 28, 2007 (continued)