2006

COLING • ACL

COLING•ACL 2006

MLRI'06
Multilingual Language Resources and Interoperability

Proceedings of the Workshop

Chairs:
Andreas Witt, Gilles Sérasset, Susan Armstrong,
Jim Breen, Ulrich Heid and Felix Sasaki

23 July 2006
Sydney, Australia

Production and Manufacturing by
BPA Digital
11 Evans St
Burwood VIC 3125
AUSTRALIA

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 1-932432-82-5

Table of Contents

Preface

In an ever-expanding information society, many language processing systems are now facing the "multilingual challenge". Language resources, such as dictionaries, thesauri and wordnets, ontologies etc., as well as annotated corpora play an important role for the development, deployment, maintenance and exploitation of language processing systems.

Much work on architectures for multilingual language resources, on recommendations of best practice for creating, representing, maintaining and upscaling such resources has been done in the 1990s, but since then, most efforts in this field have had less visibility. On the other hand, much research and development work has been done on techniques for acquisition of language data, on upper ontologies, on resource standardisation, and, last but not least, on the Semantic Web.

One of the aims of this workshop it to provide an up-to-date view on issues relating to multilingual language resources and interoperability, in terms of language description, of technology and of applications. The development and management of multilingual language resources is a long-term activity in which collaboration among researchers is essential. We hope that this workshop will gather many researchers involved in such developments and will give them the opportunity to discuss, exchange, compare their approaches and strengthen their collaborations in the field.

The impressive overall quality of the submissions (22) made the selection process quite difficult but we would like to acknowledge the dedication of our program committee who provided many useful comments to all papers. During the reviewing process we took the decision to accept only 9 papers (about 41%) in order to allow for more discussions during the workshop.

The papers address a broad range of issues related with language resources for multilingual NLP applications, covering lexicons for general and specialised language, parallel corpora, and the acquisition of data from corpora.

In particular, questions of lexical modelling and of standards for lexical resources, as well as approaches to interoperability and resource sharing in a distributed infrastructure are in focus. As multiwords are an important part of any practically usable lexical resource, two papers have been selected which deal with questions of the representation and the corpus-based acquisition of multiword items (here: collocations), from a multilingual perspective. Finally, techniques for detecting parallel texts (here: English/Japanese) and a new view on the Bible as a truly multi-lingual resource for cross-linguistic information retrieval will be discussed as examples of approaches to get access to new sources of data for the creation of language resources.

Thus, the workshop covers central aspects of resource-related research; it is structured in a way to go upstream from lexicon standardisation and sharing, over lexical modelling to the identification and the use of corpora as a source of lexical data.

The organisation of this workshop would have been impossible without the hard work of the program committee who managed to provide accurate reviews on time, on a rather tight schedule. We would also like to thank the COLING/ACL 2006 organising committee who made this workshop possible. Finally, we hope that this workshop will lead to fruitful results for all participants.

Andreas Witt, Gilles Sérasset, Susan Armstrong, Jim Breen, Ulrich Heid, Felix Sasaki

Organizers

Chairs:

Susan Armstrong, ISSCO, Université de Genève, Switzerland
Jim Breen, Monash University, Australia
Ulrich Heid, IMS-CL, University of Stuttgart, Germany
Felix Sasaki, World Wide Web Consortium (Keio Research Institute at SFC), Japan
Gilles Sérasset, GETA CLIPS-IMAG, Université Joseph Fourier, France
Andreas Witt, Bielefeld University/Eberhard-Karls-Universität Tübingen, Germany

Program Committee:

Helen Aristar-Dry, The Linguist List
Susan Armstrong, ISSCO, Université de Genève, Switzerland
Pushpak Battacharya, IIT, Mumbai, India
Christian Boitet, GETA CLIPS-IMAG, Université Joseph Fourier, France
Pierrette Bouillon, ISSCO, Université de Genève, Switzerland
Jim Breen, Monash University, Australia
Nicoletta Calzolari, CNR, Pisa, Italy
Jean Carletta, University of Edinburgh, UK
Dan Cristea, University of Iasi, Romania
Patrick Drouin, OLST, University of Montreal, Canada
Scott Farrar, University of Arizona, Tucson, USA
Ulrich Heid, IMS-CL, University of Stuttgart, Germany
Erhard Hinrichs, Eberhard-Karls-Universität Tübingen, Germany
Claus Huitfeldt, Bergen University, Norway
Phanh Huy Khan, DATIC, University of Danang, Vietnam
Nancy Ide, Vassar University, Poughkeepsie, NY, USA
Kyo Kageura, University of Tokyo, Tokyo, Japan
Chuah Choy Kim, USM, Penang, Malaisie
Anke Lüdeling, HU Berlin, Germany
Mathieu Mangeot, Université de Savoie, France
Dieter Metzing, Bielefeld University, Germany
Massimo Poesio, University of Essex, UK
Alain Polguère, OLST, University of Montreal,Canada
Andrei Popescu-belis, ISSCO, Université de Genève, Switzerland
Goutam Kumar Saha, Centre for Development of Advanced Computing, CDAC, Kolkata, India
Felix Sasaki, World Wide Web Consortium (Keio Research Institute at SFC), Japan
Thomas Schmidt, ICSI, Berkeley, USA
Gilles Sérasset, GETA CLIPS-IMAG, Université Joseph Fourier, France
Gary Simons, SIL International, USA
Virach Sornlertlamvanich, Thai Computational Linguistics Laboratory, NICT, Thailand
C.M. Sperberg-McQueen, MIT Boston and W3C, USA
Manfred Stede, Potsdam University, Germany

Koichi Takeuchi, Okayama University, Japan
Dan Tufis RACAI, Uni Bucharest, Romania
Jun'ichi Tsujii, University of Tokyo, Japan
Takehito Utsuro, Kyoto University, Japan
Andreas Witt, Bielefeld University/Eberhard-Karls-Universität Tübingen, Germany
Michael Zock, LIF-CNRS, Marseille, France

Additional Reviewer:

Laurent Besacier, GEOD CLIPS-IMAG, Université Joseph Fourier, France

Workshop Program

Sunday, 23 July 2006

8:45–9:00 Registration

9:10–9:20 Opening Remarks

9:20–9:55 *Lexical Markup Framework (LMF) for NLP Multilingual Resources*
Gil Francopoulo, Nuria Bel, Monte George, Nicoletta Calzolari, Monica Monachini, Mandy Pet and Claudia Soria

9:55–10:30 *The Role of Lexical Resources in CJK Natural Language Processing*
Jack Halpern

10:30–11:00 Coffee break

11:00–11:35 *Towards Agent-based Cross-Lingual Interoperability of Distributed Lexical Resources*
Claudia Soria, Maurizio Tesconi, Andrea Marchetti, Francesca Bertagna, Monica Monachini, Chu-Ren Huang and Nicoletta Calzolari

11:35–12:10 *The LexALP Information System: Term Bank and Corpus for Multilingual Legal Terminology Consolidated*
Verena Lyding, Elena Chiocchetti, Gilles Sérasset and Francis Brunet-Manquat

12:10–13:45 Lunch break

13:45–14:20 *The Development of a Multilingual Collocation Dictionary*
Sylviane Cardey, Rosita Chan and Peter Greenfield

14:20–14:55 *Multilingual Collocation Extraction: Issues and Solutions*
Violeta Seretan and Eric Wehrli

14:55–15:30 *Structural Properties of Lexical Systems: Monolingual and Multilingual Perspectives*
Alain Polguère

15:30–16:00 Coffee break

16:00–16:35 *A Fast and Accurate Method for Detecting English-Japanese Parallel Texts*
Ken’ichi Fukushima, Kenjiro Taura and Takashi Chikayama

16:35–17:10 *Evaluation of the Bible as a Resource for Cross-Language Information Retrieval*
Peter A. Chew, Steve J. Verzi, Travis L. Bauer and Jonathan T. McClain

17:10–17:30 Closing remarks