ACL-IJCNLP 2009

MWE 2009

2009 Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation, Applications

Proceedings of the Workshop

6 August 2009
Suntec, Singapore

Production and Manufacturing by
World Scientific Publishing Co Pte Ltd
5 Toh Tuck Link
Singapore 596224

©2009 The Association for Computational Linguistics
and The Asian Federation of Natural Language Processing

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 978-1-932432-60-2 / 1-932432-60-4

Introduction

The ACL 2009 Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications (MWE'09) took place on August 6, 2009 in Singapore, immediately following the annual meeting of the Association for Computational Linguistics (ACL). This is the fifth time this workshop has been held in conjunction with ACL, following the meetings in 2003, 2004, 2006, and 2007.

The workshop focused on Multi-Word Expressions (MWEs), which represent an indispensable part of natural languages and appear steadily on a daily basis, both novel and already existing but paraphrased, which makes them important for many natural language applications. Unfortunately, while easily mastered by native speakers, MWEs are often non-compositional, which poses a major challenge for both foreign language learners and automatic analysis.

The growing interest in MWEs in the NLP community has led to many specialized workshops held every year since 2001 in conjunction with ACL, EACL and LREC; there have been also two recent special issues on MWEs published by leading journals: the International Journal of Language Resources and Evaluation, and the Journal of Computer Speech and Language.

As a result of the overall progress in the field, the time has come to move from basic preliminary research to actual applications in real-world NLP tasks. Thus, in MWE'09, we were interested in the overall process of dealing with MWEs, asking for original research on the following four fundamental topics:

Identification. Identifying MWEs in free text is a very challenging problem. Due to the variability of expression, it does not suffice to collect and use a static list of known MWEs; complex rules and machine learning are typically needed as well.

Interpretation. Semantically interpreting MWEs is a central issue. For some kinds of MWEs, e.g., noun compounds, it could mean specifying their semantics using a static inventory of semantic relations, e.g., WordNet-derived. In other cases, MWE's semantics could be expressible by a suitable paraphrase.

Disambiguation. Most MWEs are ambiguous in various ways. A typical disambiguation task is to determine whether an MWE is used non-compositionally (i.e., figuratively) or compositionally (i.e., literally) in a particular context.

Applications. Identifying MWEs in context and understanding their syntax and semantics is important for many natural language applications, including but not limited to question answering, machine translation, information retrieval, information extraction, and textual entailment. Still, despite the growing research interest, there are not enough successful applications in real NLP problems, which we believe is the key for the advancement of the field.

Of course, the above topics largely overlap. For example, identification can require disambiguating between literal and idiomatic uses since MWEs are typically required to be non-compositional by definition. Similarly, interpreting three-word noun compounds like *morning flight ticket* and *plastic water bottle* requires disambiguation between a left and a right syntactic structure, while interpreting two-word compounds like *English teacher* requires disambiguating between (a) 'teacher who teaches English' and (b) 'teacher coming from England (who could teach any subject, e.g., math)'.

We received 18 submissions, and, given our limited capacity as a one-day workshop, we were only able to accept 9 full papers for oral presentation, an acceptance rate of 50%.

We would like to thank the members of the Program Committee for their timely reviews. We would also like to thank the authors for their valuable contributions.

Dimitra Anastasiou, Chikara Hashimoto, Preslav Nakov, and Su Nam Kim
Co-Organizers

Organizers:

Dimitra Anastasiou, Localisation Research Centre, Limerick University, Ireland
Chikara Hashimoto, National Institute of Information and Communications Technology, Japan
Preslav Nakov, National University of Singapore, Singapore
Su Nam Kim, University of Melbourne, Australia

Program Committee:

Iñaki Alegria, University of the Basque Country (Spain)
Timothy Baldwin, University of Melbourne (Australia)
Colin Bannard, Max Planck Institute (Germany)
Francis Bond, National Institute of Information and Communications Technology (Japan)
Gaël Dias, Beira Interior University (Portugal)
Ulrich Heid, Stuttgart University (Germany)
Stefan Evert, University of Osnabrück (Germany)
Afsaneh Fazly, University of Toronto (Canada)
Nicole Grégoire, University of Utrecht (The Netherlands)
Roxana Girju, University of Illinois at Urbana-Champaign (USA)
Kyo Kageura, University of Tokyo (Japan)
Brigitte Krenn, Austrian Research Institute for Artificial Intelligence (Austria)
Éric Laporte, University of Marne-la-Vallée (France)
Rosamund Moon, University of Birmingham (UK)
Diana McCarthy, University of Sussex (UK)
Jan Odijk, University of Utrecht (The Netherlands)
Stephan Oepen, University of Oslo (Norway)
Darren Pearce, London Knowledge Lab (UK)
Pavel Pecina, Charles University (Czech Republic)
Scott Piao, University of Manchester (UK)
Violeta Seretan, University of Geneva (Switzerland)
Stan Szpakowicz, University of Ottawa (Canada)
Beata Trawinski, University of Tübingen (Germany)
Peter Turney, National Research Council of Canada (Canada)
Kiyoko Uchiyama, Keio University (Japan)
Begoña Villada Moirón, University of Groningen (The Netherlands)
Aline Villavicencio, Federal University of Rio Grande do Sul (Brazil)

Table of Contents

Workshop Program

Friday, August 6, 2009

8:30–8:45 Welcome and Introduction to the Workshop

Session 1 (08:45–10:00): MWE Identification and Disambiguation

08:45–09:10 *Statistically-Driven Alignment-Based Multiword Expression Identification for Technical Domains*
Helena Caseli, Aline Villavicencio, André Machado and Maria José Finatto

09:10–09:35 *Re-examining Automatic Keyphrase Extraction Approaches in Scientific Articles*
Su Nam Kim and Min-Yen Kan

09:35–10:00 *Verb Noun Construction MWE Token Classification*
Mona Diab and Pravin Bhutada

10:00-10:30 BREAK

Session 2 (10:30–12:10): Identification, Interpretation, and Disambiguation

10:30–10:55 *Exploiting Translational Correspondences for Pattern-Independent MWE Identification*
Sina Zarrieß and Jonas Kuhn

10:55–11:20 *A re-examination of lexical association measures*
Hung Huu Hoang, Su Nam Kim and Min-Yen Kan

11:20–11:45 *Mining Complex Predicates In Hindi Using A Parallel Hindi-English Corpus*
R. Mahesh K. Sinha

11:45-13:50 LUNCH

Session 3 (13:50–15:30): Applications

13:50–14:15 *Improving Statistical Machine Translation Using Domain Bilingual Multiword Expressions*
Zhixiang Ren, Yajuan Lü, Jie Cao, Qun Liu and Yun Huang

14:15–14:40 *Bottom-up Named Entity Recognition using Two-stage Machine Learning Method*
Hirotaka Funayama, Tomohide Shibata and Sadao Kurohashi

14:40–15:05 *Abbreviation Generation for Japanese Multi-Word Expressions*
Hiromi Wakaki, Hiroko Fujii, Masaru Suzuki, Mika Fukui and Kazuo Sumita

15:05-15:30 Discussion of Sessions 1, 2, 3 (Creating an Agenda for the general discussion)

15:30-16:00 BREAK

16:00-17:00 General Discussion

17:00-17:15 Closing Remarks