Coling 2010

23rd International Conference on Computational Linguistics

Proceedings of the
Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)

28 August 2010
Beijing International Convention Center

Produced by
Chinese Information Processing Society of China

To order the CD of Coling 2010 and its Workshop Proceedings, please contact:

Chinese Information Processing Society of China
No.4, Southern Fourth Street
Haidian District, Beijing, 100190
China
Tel: +86-010-62562916
Fax: +86-010-62562916
cips@iscas.ac.cn

Introduction

The COLING 2010 Workshop on *Multiword Expressions: from Theory to Applications* (MWE 2010) took place on August 28, 2010 in Beijing, China, following the 23rd International Conference on Computational Linguistics (COLING 2010). The workshop has been held every year since 2003 in conjunction with ACL, EACL and LREC; this is the first time that it has been co-located with COLING.

Multiword Expressions (MWEs) are a ubiquitous component of natural languages and appear steadily on a daily basis, both in specialized and in general-purpose communication. While easily mastered by native speakers, their interpretation poses a major challenge for automated analysis due to their flexible and heterogeneous nature. Therefore, the automated processing of MWEs is desirable for any natural language application that involves some degree of semantic interpretation, e.g., Machine Translation, Information Extraction, and Question Answering.

In spite of the recent advances in the field, there is a wide range of open problems that prevent MWE treatment techniques from full integration in current NLP systems. In MWE'2010, we were interested in major challenges in the overall process of MWE treatment. We thus asked for original research related but not limited to the following topics:

- **MWE resources:** Although underused in most current state-of-the-art approaches, resources are key for developing real-world applications capable of interpreting MWEs. We thus encouraged submissions describing the process of building MWE resources, constructed both manually and automatically from text corpora; we were also interested in assessing the usability of such resources in various MWE tasks.
- **Hybrid approaches:** We further invited research on integrating heterogeneous MWE treatment techniques and resources in NLP applications. Such hybrid approaches can aim, for example, at the combination of results from symbolic and statistical approaches, at the fusion of manually built and automatically extracted resources, or at the design of language learning techniques.
- **Domain adaptation:** Real-world NLP applications need to be robust to deal with texts coming from different domains. Thus, its is important to assess the performance of MWE methods across domains or describing domain adaptation techniques for MWEs.
- **Multilingualism:** Parallel and comparable corpora are gaining popularity as a resource for automatic MWE discovery and treatment. We were thus interested in the integration of MWE processing in multilingual applications such as machine translation and multilingual information retrieval, as well as in porting existing monolingual MWE approaches to new languages.

We received 18 submissions, and, given our limited capacity as a one-day workshop, we were only able to accept eight full papers for oral presentation: an acceptance rate of 44%. We further accepted four papers as posters. The regular papers were distributed in three sessions: Lexical Representation, Identification and Extraction, and Applications. The workshop also featured two invited talks, by Kyo Kageura and by Aravind K. Joshi, and a panel discussion.

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.

Éric Laporte, Preslav Nakov, Carlos Ramisch, Aline Villavicencio
Co-Organizers

Organizers:

Éric Laporte, Université Paris-Est
Preslav Nakov, National University of Singapore
Carlos Ramisch, University of Grenoble and Federal University of Rio Grande do Sul
Aline Villavicencio, Federal University of Rio Grande do Sul

Program Committee:

Iñaki Alegria, University of the Basque Country
Dimitra Anastasiou, Limerick University
Timothy Baldwin, University of Melbourne
Colin Bannard, University of Texas at Austin
Francis Bond, Nanyang Technological University
Paul Cook, University of Toronto
Béatrice Daille, Nantes University
Gaël Dias, Beira Interior University
Stefan Evert, University of Osnabrück
Roxana Girju, University of Illinois at Urbana-Champaign
Nicole Grégoire, University of Utrecht
Chikara Hashimoto, National Institute of Information and Communications Technology
Marti Hearst, University of California at Berkeley
Kyo Kageura, University of Tokyo
Min-Yen Kan, National University of Singapore
Adam Kilgarriff, Lexical Computing Ltd
Su Nam Kim, University of Melbourne
Anna Korhonen, University of Cambridge
Zornitsa Kozareva, University of Southern California
Brigitte Krenn, Austrian Research Institute for Artificial Intelligence
Cvetana Krstev, University of Belgrade
Rosamund Moon, University of Birmingham
Diarmuid Ó Séaghdha, University of Cambridge
Jan Odijk, University of Utrecht
Stephan Oepen, University of Oslo
Darren Pearce-Lazard, University of Sussex
Pavel Pecina, Dublin City University
Scott Piao, Lancaster University
Thierry Poibeau, CNRS and École Normale Supérieure
Elisabete Ranchhod, University of Lisbon
Barbara Rosario, Intel Labs
Violeta Seretan, University of Geneva
Stan Szpakowicz, University of Ottawa

Beata Trawinski, University of Vienna
Vivian Tsang, Bloorview Research Institute
Kyioko Uchiyama, National Institute of Informatics
Ruben Urizar, University of the Basque Country
Tony Veale, University College Dublin

Invited Speakers:

Kyo Kageura, University of Tokyo
Aravind K. Joshi, University of Pennsylvania

Table of Contents

Workshop Program

Saturday, August 28, 2010

08:30–08:40 **Welcome**

08:40–09:40 **Invited Talk**
Being Theoretical is Being Practical: Multiword Units and Terminological Structure Revitalised
Kyo Kageura, University of Tokyo

Session I: Lexical Representation
Chair: Pavel Pecina

09:40–10:05 *Computational Lexicography of Multi-Word Units: How Efficient Can It Be?*
Filip Graliński, Agata Savary, Monika Czerepowicka and Filip Makowiecki

10:05–10:30 *Construction of a Chinese Idiom Knowledge Base and Its Applications*
Lei Wang and Shiwen Yu

10:30–11:00 **Break**

Session II: Identification and Extraction
Chair: Aline Villavicencio

11:00–11:25 *Automatic Extraction of Arabic Multiword Expressions*
Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina and Josef van Genabith

11:25–11:50 *Sentence Analysis and Collocation Identification*
Eric Wehrli, Violeta Seretan and Luka Nerima

11:50–12:15 *Automatic Extraction of Complex Predicates in Bengali*
Dipankar Das, Santanu Pal, Tapabrata Mondal, Tanmoy Chakraborty and Sivaji Bandyopadhyay

12:15–13:50 **Lunch**

Session III: Applications
Chair: Eric Wehrli

13:50–14:15 *Handling Named Entities and Compound Verbs in Phrase-Based Statistical Machine Translation*
Santanu Pal, Sudip Kumar Naskar, Pavel Pecina, Sivaji Bandyopadhyay and Andy Way

14:15–14:40 *Application of the Tightness Continuum Measure to Chinese Information Retrieval*
Ying Xu, Randy Goebel, Christoph Ringlstetter and Grzegorz Kondrak

14:40–15:05 *Standardizing Complex Functional Expressions in Japanese Predicates: Applying Theoretically-Based Paraphrasing Rules*
Tomoko Izumi, Kenji Imamura, Genichiro Kikui and Satoshi Sato

Saturday, August 28, 2010 (continued)

15:05–15:30 **Poster Session**
Chair: Carlos Ramisch

Identification of Reduplication in Bengali Corpus and their Semantic Analysis: A Rule Based Approach
Tanmoy Chakraborty and Sivaji Bandyopadhyay
Contrastive Filtering of Domain-Specific Multi-Word Terms from Different Types of Corpora
Francesca Bonin, Felice Dell'Orletta, Giulia Venturi and Simonetta Montemagni
A Hybrid Approach for Functional Expression Identification in a Japanese Reading Assistant
Gregory Hazelbeck and Hiroaki Saito
An Efficient, Generic Approach to Extracting Multi-Word Expressions from Dependency Trees
Scott Martens and Vincent Vandeghinste

15:30–16:00 **Break**

16:00–17:00 **Invited Talk**
Multiword Expressions as Discourse Relation Markers (DRMs)
Aravind Joshi, University of Pennsylvania

17:00–17:50 **Panel: Multiword Expressions – from Theory to Applications**
Moderator: Aline Villavicencio

Mona Diab, Columbia University
Valia Kordoni, Saarland University
Hans Uszkoreit, Saarland University

17:50–18:00 **Closing Remarks**