ACL 2007

Proceedings of the Workshop on A Broader Perspective on Multiword Expressions

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

This volume contains the papers accepted for presentation at the workshop *A Broader Perspective on Multiword Expressions*. The workshop is endorsed by the Association for Computational Linguistics Special Interest Group on the Lexicon (SIGLEX) and is held in conjunction with the ACL 2007 Conference on June 28th, 2007 in Prague, Czech Republic.

In recent years, the NLP community has increasingly become aware of the problems that multiword expressions (MWEs) pose. A considerable amount of research has been conducted in this area, some within large research projects dedicated to MWEs. Although progress has been made especially in the area of multiword extraction, a number of fundamental questions remain unanswered. The goal of the workshop is to address some of these questions with oral and poster presentations, as well as general discussion period at the end of the workshop. In particular, we want to focus on the following topics:

- Is it sufficient to use purely statistical methods for the extraction of MWEs from corpora, or is it necessary to harness human knowledge and linguistic insights?
- To what extent can definitions and extraction procedures be generalised to other languages, other text types and other types of MWEs?
- What properties should be specified for MWEs or subtypes of MWEs in the lexicon? And can we detect these properties automatically with sufficient accuracy?
- What role do the semantics of MWEs play in NLP applications and can they be determined automatically from large corpora?

We received 23 submissions in total. Each submission was reviewed by at least two members of the program committee, who did not only give an overall verdict but also provided detailed comments to the authors. Due to the large number of interesting papers we had received and the fact that the workshop is only half-day, we decided on an unusual format including a poster session slot. This allowed us to accept ten papers for presentation at the workshop, four oral and six poster presentations. The poster session offers an opportunity to exhibit a wider range of approaches and points of view than would otherwise have been possible, and we hope it will thus initiate a lively and fruitful discussion period at the end of the workshop.

We would like to thank all the authors for submitting their research and the members of the program committee for their careful reviews and useful suggestions to the authors. We would also like to thank the ACL 2007 organising committee that made this workshop possible and SIGLEX for its endorsement.

Finally, we hope that this workshop will provide plentiful and tasty food for thought to all participants as well as readers of its proceedings.

Nicole Grégoire
Stefan Evert
Su Nam Kim

Organizers

Chairs:

Nicole Grégoire, University of Utrecht (The Netherlands)
Stefan Evert, University of Osnabrueck (Germany)
Su Nam Kim, University of Melbourne (Australia)

Program Committee:

Iñaki Alegria, University of the Basque Country (Spain)
Timothy Baldwin, Stanford University (USA); University of Melbourne (Australia)
Francis Bond, NTT Communication Science Laboratories (Japan)
Beatrice Daille, Nantes University (France)
Gael Dias, Beira Interior University (Portugal)
Kyo Kageura, University of Tokyo (Japan)
Anna Korhonen, University of Cambridge (UK)
Rosamund Moon, University of Birmingham (UK)
Diana McCarthy, University of Sussex (UK)
Eric Laporte, University of Marne-la-Vallee (France)
Preslav Nakov, University of California, Berkeley (USA)
Jan Odijk, University of Utrecht (The Netherlands)
Stephan Oepen, Stanford University (USA); University of Oslo (Norway)
Darren Pearce, University of Sussex (UK)
Scott Piao, University of Manchester (UK)
Violeta Seretan, University of Geneva (Switzerland)
Suzanne Stevenson, University of Toronto (Canada)
Beata Trawinski, University of Tuebingen (Germany)
Vivian Tsang, University of Toronto (Canada) Kiyoko Uchiyama, Keio University (Japan)
Ruben Urizar, University of the Basque Country (Spain)
Begoña Villada Moirón, University of Groningen (The Netherlands)
Aline Villavicencio, Federal University of Rio Grande do Sul (Brazil)

Table of Contents

Conference Program

Thursday, 28 June 2007

09:00–09:10 Opening remarks

09:10–10:50 Oral presentations

A Measure of Syntactic Flexibility for Automatically Identifying Multiword Expressions in Corpora
Colin Bannard

Distinguishing Subtypes of Multiword Expressions Using Linguistically-Motivated Statistical Measures
Afsaneh Fazly and Suzanne Stevenson

Design and Implementation of a Lexicon of Dutch Multiword Expressions
Nicole Grégoire

Semantics-based Multiword Expression Extraction
Tim Van de Cruys and Begoña Villada Moirón

10:50–11:20 Coffee break

11:20–11:40 Poster introduction (6x3 minutes)

11:40–12:30 Poster session

Spanish Adverbial Frozen Expressions
Dolors Català and Jorge Baptista

Pulling their Weight: Exploiting Syntactic Forms for the Automatic Identification of Idiomatic Expressions in Context
Paul Cook, Afsaneh Fazly and Suzanne Stevenson

Pauses as an Indicator of Psycholinguistically Valid Multi-Word Expressions (MWEs)?
Irina Dahlmann and Svenja Adolphs

Co-occurrence Contexts for Noun Compound Interpretation
Diarmuid Ó Séaghdha and Ann Copestake

Thursday, 28 June 2007 (continued)

12:30–13:00 Discussion and closing