ACL HLT 2011

Workshop on Multiword Expressions:
from Parsing and Generation to the Real World
MWE 2011

Proceedings of the Workshop

23 June, 2011
Portland, Oregon, USA

Production and Manufacturing by
Omnipress, Inc.
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

ISBN 978-1-932432-97-8

Introduction

The ACL 2011 Workshop on *Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011)* took place on June 23, 2011 in Portland, Oregon, USA, in conjunction to the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011). The workshop has been held every year since 2003 in conjunction with ACL, EACL, COLING and LREC.

Multiword Expressions (MWEs) range over linguistic constructions such as idioms (a frog in the throat, kill some time), fixed phrases (per se, by and large, rock'n roll), noun compounds (telephone booth, cable car), compound verbs (give a presentation, go by [a name]), etc. While easily mastered by native speakers, their interpretation poses a major challenge for computational systems, due to their flexible and heterogeneous nature. Surprisingly enough, MWEs are not nearly as frequent in NLP resources (dictionaries, grammars) as they are in real-word text, where they have been reported to account for over 70% of the terms in a domain. Thus, MWEs are a key issue and a current weakness for tasks like Natural Language Parsing (NLP) and Generation (NLG), as well as real-life applications such as Machine Translation.

MWE 2011 is the 8th event in the series, and the time has come to move from basic preliminary research and theoretical results to actual applications in real-world NLP tasks. Therefore, following further the trend of previous MWE workshops, we have now turned our focus towards MWEs on NLP applications, specifically towards Parsing and Generation of MWEs, as there is a wide range of open problems that prevent MWE treatment techniques to be fully integrated in current NLP systems. We have thus asked our contributors for original research related (but not limited) to the following topics:

- **Lexical representations**: In spite of several proposals for MWE representation ranging along the continuum from words-with-spaces to compositional approaches connecting lexicon and grammar, to date, it remains unclear how MWEs should be represented in electronic dictionaries, thesauri and grammars. New methodologies that take into account the type of MWE and its properties are needed for efficiently handling manually and/or automatically acquired expressions in NLP systems. Moreover, strategies are also needed to represent deep attributes and semantic properties for these multiword entries.
- **Task and Application-oriented evaluation**: Evaluation is a crucial aspect for MWE research. Various evaluation techniques have been proposed, from manual inspection of top-n candidates to classic precision/recall measures. However, to get a clear indication of the effect of incorporating a treatment of MWEs in a particular context, task and application-oriented evaluations are needed. We have thus called for submissions that study the impact of MWE handling in the context of Parsing, Generation, Information Extraction, Machine Translation, Summarization, etc.
- **Type-dependent analysis**: While there is no unique definition or classification of MWEs, most researchers agree on some major classes such as named entities, collocations, multiword terminology and verbal expressions. These, though, are very heterogeneous in terms of syntactic and semantic properties, and should thus be treated differently by applications. Type-dependent analyses could shed some light on the best methodologies to integrate MWE knowledge in our analysis and generation systems.

- **MWE engineering**: Where do MWEs go after being extracted? Do they belong to the lexicon and/or to the grammar? In the pipeline of linguistic analysis and/or generation, where should we insert MWEs? And even more important: HOW? Because all the effort put in automatic MWE extraction will not be useful if we do not know how to employ these rich resources in our real-life NLP applications!

This year, we had three different submission types: long, short and demonstration papers. We received a total of 31 submissions, from which 16 were long papers, 9 were short papers and 6 were demo papers. Given our limited capacity as a one-day workshop, we were only able to accept 6 long papers for oral presentation and 4 long papers as posters: an acceptance rate of 62.5%. We further accepted 4 short papers for oral presentation and 2 short papers as posters (67% acceptance), as well as 5 out of the 6 proposed demonstrations. The oral presentations were distributed in three sessions: Short Papers, Identification and Representation, and Tasks and Applications. The workshop also featured two invited talks, by Timothy Baldwin and by Kenneth Church, and a panel discussion.

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

Valia Kordoni, Carlos Ramisch, Aline Villavicencio
Co-Organizers

Organizers:

Valia Kordoni, DFKI GmbH and Saarland University, Germany
Carlos Ramisch, University of Grenoble, France and Federal University of Rio Grande do Sul, Brazil
Aline Villavicencio, Federal University of Rio Grande do Sul, Brazil

Consulting Body:

Su Nam Kim, University of Melbourne, Australia
Preslav Nakov, National University of Singapore, Singapore

Program Committee:

Iñaki Alegria, University of the Basque Country, Spain
Dimitra Anastasiou, University of Bremen, Germany
Timothy Baldwin, University of Melbourne, Australia
Srinivas Bangalore, AT&T Labs-Research, USA
Francis Bond, Nanyang Technological University, Singapore
Aoife Cahill, IMS University of Stuttgart, Germany
Paul Cook, University of Toronto, Canada
Béatrice Daille, Nantes University, France
Mona Diab, Columbia University, USA
Gaël Dias, Beira Interior University, Portugal
Stefan Evert, University of Osnabrueck, Germany
Roxana Girju, University of Illinois at Urbana-Champaign, USA
Chikara Hashimoto, National Institute of Information and Communications Technology, Japan
Ulrich Heid, Stuttgart University, Germany
Kyo Kageura, University of Tokyo, Japan
Adam Kilgarriff, Lexical Computing Ltd., UK
Ioannis Korkontzelos, University of Manchester, UK
Zornitsa Kozareva, University of Southern California, USA
Brigitte Krenn, Austrian Research Institute for Artificial Intelligence, Austria
Takuya Matsuzaki, University of Tokyo, Japan
Diana McCarthy, Lexical Computing Ltd., UK
Yusuke Miyao, National Institute of Informatics, Japan
Rosamund Moon, University of Birmingham, UK
Diarmuid Ó Séaghdha, University of Cambridge, UK
Jan Odijk, University of Utrecht, The Netherlands
Pavel Pecina, Dublin City University, Ireland
Scott Piao, Lancaster University, UK
Thierry Poibeau, CNRS and École Normale Supérieure, France

Elisabete Ranchhod, University of Lisbon, Portugal
Barbara Rosario, Intel Labs, USA
Agata Savary, Université François Rabelais Tours, France
Violeta Seretan, University of Edinburgh, UK
Ekaterina Shutova, University of Cambridge, UK
Suzanne Stevenson, University of Toronto, Canada
Sara Stymne, Linköping University, Sweden
Stan Szpakowicz, University of Ottawa, Canada
Beata Trawinski, University of Vienna, Austria
Vivian Tsang, Bloorview Research Institute, Canada
Kyioko Uchiyama, National Institute of Informatics, Japan
Ruben Urizar, University of the Basque Country, Spain
Gertjan van Noord, University of Groningen, The Netherlands
Tony Veale, University College Dublin, Ireland
Begoña Villada Moirón, RightNow, The Netherlands
Yi Zhang, DFKI GmbH and Saarland University, Germany

Invited Speakers:

Timothy Baldwin, University of Melbourne, Australia
Kenneth Church, Johns Hopkins University, USA

Table of Contents

Workshop Program

Thursday, June 23, 2011

08:15–08:30 Welcome

08:30–09:30 **Invited talk**
MWEs and Topic Modelling: Enhancing Machine Learning with Linguistics
Timothy Baldwin

Session I - Short Papers

09:30–09:45 *Automatic Extraction of NV Expressions in Basque: Basic Issues on Cooccurrence Techniques*
Antton Gurrutxaga and Iñaki Alegria

09:45–10:00 *Semantic Clustering: an Attempt to Identify Multiword Expressions in Bengali*
Tanmoy Chakraborty, Dipankar Das and Sivaji Bandyopadhyay

10:00–10:15 *Decreasing Lexical Data Sparsity in Statistical Syntactic Parsing - Experiments with Named Entities*
Deirdre Hogan, Jennifer Foster and Josef van Genabith

10:15–10:30 *Detecting Multi-Word Expressions Improves Word Sense Disambiguation*
Mark Finlayson and Nidhi Kulkarni

10:30–11:00 MORNING BREAK

Session II - Identification and Representation

11:00–11:25 *Tree-Rewriting Models of Multi-Word Expressions*
William Schuler and Aravind Joshi

11:25–11:50 *Learning English Light Verb Constructions: Contextual or Statistical*
Yuancheng Tu and Dan Roth

11:50–12:15 *Two Types of Korean Light Verb Constructions in a Typed Feature Structure Grammar*
Juwon Lee

12:15–13:50 LUNCH BREAK

Session III - Tasks and Applications

13:50–14:15 *MWU-Aware Part-of-Speech Tagging with a CRF Model and Lexical Resources*
Matthieu Constant and Anthony Sigogne

14:15–14:40 *The Web is not a PERSON, Berners-Lee is not an ORGANIZATION, and African-Americans are not LOCATIONS: An Analysis of the Performance of Named-Entity Recognition*
Robert Krovetz, Paul Deane and Nitin Madnani

14:40–15:05 *A Machine Learning Approach to Relational Noun Mining in German*
Berthold Crysmann

Thursday, June 23, 2011 (continued)

15:05–15:30 **Poster and Demo Session**
Long Papers

Identifying and Analyzing Brazilian Portuguese Complex Predicates
Magali Sanches Duran, Carlos Ramisch, Sandra Maria Aluísio and Aline Villavicencio

An N-gram Frequency Database Reference to Handle MWE Extraction in NLP Applications
Patrick Watrin and Thomas François

Extracting Transfer Rules for Multiword Expressions from Parallel Corpora
Petter Haugereid and Francis Bond

Identification and Treatment of Multiword Expressions Applied to Information Retrieval
Otavio Acosta, Aline Villavicencio and Viviane Moreira

Short Papers

Stepwise Mining of Multi-Word Expressions in Hindi
Rai Mahesh Sinha

Detecting Noun Compounds and Light Verb Constructions: a Contrastive Study
Veronika Vincze, István Nagy T. and Gábor Berend

Demo Papers

jMWE: A Java Toolkit for Detecting Multi-Word Expressions
Nidhi Kulkarni and Mark Finlayson

FipsCoView: On-line Visualisation of Collocations Extracted from Multilingual Parallel Corpora
Violeta Seretan and Eric Wehrli

The StringNet Lexico-Grammatical Knowledgebase and its Applications
David Wible and Nai-Lung Tsao

The Ngram Statistics Package (Text::NSP) : A Flexible Tool for Identifying Ngrams, Collocations, and Word Associations
Ted Pedersen, Satanjeev Banerjee, Bridget McInnes, Saiyam Kohli, Mahesh Joshi and Ying Liu

Fast and Flexible MWE Candidate Generation with the mwetoolkit
Vitor De Araujo, Carlos Ramisch and Aline Villavicencio

15:30–16:00 AFTERNOON BREAK

Invited talk
16:00–17:00 *How Many Multiword Expressions do People Know?*
Kenneth Church

17:00–18:00 **Panel: Toward a Special Interest Group for MWEs**