ACL 2016

The 54th Annual Meeting of the Association for Computational Linguistics

Proceedings of the 12th Workshop on Multiword Expressions (MWE'2016)

August 11, 2016
Berlin, Germany

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 978-1-945626-06-7

Introduction

The 12th Workshop on Multiword Expressions (MWE'2016) took place on August 11, 2016 in Berlin, Germany, in conjunction with the 54th Annual Meeting of the Association for Computational Linguistics (ACL'2016) and was endorsed by the Special Interest Group on the Lexicon of the Association for Computational Linguistics (SIGLEX), as well as by the SIGLEX Section dedicated to the study and research of Multiword Expressions (SIGLEX-MWE).

The workshop has been held almost every year since 2003, in conjunction with ACL, EACL, NAACL, COLING, and LREC. It is the main venue of the field for interaction, sharing of resources and tools and collaboration efforts for advancing the computational treatment of Multiword Expressions (MWEs), attracting the attention of an ever-growing community from all around the world working on a variety of languages and MWE types.

MWEs include idioms (*storm in a teacup*, *sweep under the rug*), fixed phrases (*in vitro*, *by and large*), noun compounds (*olive oil*, *laser printer*), compound verbs (*take a nap*, *bring about*), among others. These, while easily mastered by native speakers, are a key issue and a current weakness for natural language parsing and generation, as well as for real-life applications that require some degree of semantic interpretation, such as machine translation, just to name a prominent one among many. However, thanks to the joint efforts of researchers from several fields working on MWEs, significant progress has been made in recent years, especially concerning the construction of large-scale language resources. For instance, there is a large number of recent papers that focus on the acquisition of MWEs from corpora, and others that describe a variety of techniques to find paraphrases for MWEs. Current methods use a plethora of tools such as association measures, machine learning, syntactic patterns, web queries, etc.

In the call for papers, we solicited submissions about major challenges in the overall process of MWE treatment, both from a theoretical and a computational viewpoint, focusing on original research related (but not limited) to the following topics:

- Lexicon-grammar interface for MWEs
- Parsing techniques for MWEs
- Hybrid parsing of MWEs
- Annotating MWEs in treebanks
- MWEs in Machine Translation and Translation Technology
- Manually and automatically constructed resources
- Representation of MWEs in dictionaries and ontologies
- MWEs and user interaction
- Multilingual acquisition
- Multilingualism and MWE processing
- Models of first and second language acquisition of MWEs
- Crosslinguistic studies on MWEs
- The role of MWEs in the domain adaptation of parsers
- Integration of MWEs into NLP applications
- Evaluation of MWE treatment techniques
- Lexical, syntactic or semantic aspects of MWEs

Submission modalities included long papers and short papers. From a total of 49 submissions, we accepted 4 long papers for oral presentation. We further accepted 5 short papers for oral presentation and another 8 short papers as posters. Thus the total number of accepted papers is 18, or an overall acceptance rate of 37%.

Acknowledgements

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

Valia Kordoni, Kostadin Cholakov, Markus Egg, Stella Markantonatou, Preslav Nakov
Co-Organizers

Organizers:

Valia Kordoni, Humboldt Universität zu Berlin (Germany)
Kostadin Cholakov, Humboldt Universität zu Berlin (Germany)
Markus Egg, Humboldt Universität zu Berlin (Germany)
Stella Markantonatou, Institute for Language and Speech Processing (ILSP) - Athena Research Center (Greece)
Preslav Nakov, Qatar Computing Research Institute, HBKU (Qatar)

Program Committee:

Dimitra Anastasiou, LIST-Luxembourg Institute of Science and Technology (Luxembourg)
Tim Baldwin, The University of Melbourne (Australia)
Núria Bel, Pompeu Fabra University (Spain)
Lars Borin, University of Gothenburg (Sweden)
Jill Burstein, ETS (USA)
Aoife Cahill, ETS (USA)
Paul Cook, University of New Brunswick (Canada)
Anastasia Christofidou, Academy of Athens/National and Kapodistrian University of Athens (Greece)
Béatrice Daille, Nantes University (France)
Joaquim Ferreira da Silva, New University of Lisbon (Portugal)
Aggeliki Fotopoulou, Institute for Language and Speech Processing / Athena Research Center (Greece)
Voula Gotsoulia, National and Kapodistrian University of Athens (Greece)
Chikara Hashimoto, National Institute of Information and Communications Technology (Japan)
Kyo Kageura, University of Tokyo (Japan)
Dimitrios Kokkinakis, University of Gothenburg (Sweden)
Ioannis Korkontzelos, University of Manchester (UK)
Takuya Matsuzaki, Nagoya University (Japan)
Yusuke Miyao, National Institute of Informatics (Japan)
Joakim Nivre, University of Uppsala (Sweden)
Diarmuid Ó Séaghdha, University of Cambridge and VocalIQ (UK)
Haris Papageorgiou, Institute for Language and Speech Processing/Athena Research Center (Greece)
Yannick Parmentier, University of Orleans (France)
Pavel Pecina, Charles University in Prague (Czech Republic)
Scott Piao, Lancaster University (UK)
Barbara Plank, University of Groningen (The Netherlands)
Maja Popović, Humboldt Universität zu Berlin (Germany)
Prokopidis Prokopis, Institute for Language and Speech Processing/Athena Research Center (Greece)
Carlos Ramisch, Aix-Marseille University (France)
Martin Riedl, University of Darmstadt (Germany)
Will Roberts, Humboldt Universität zu Berlin (Germany)
Agata Savary, Université François Rabelais Tours (France)
Aline Villavicencio, Federal University of Rio Grande do Sul (Brazil)
Veronika Vincze, Hungarian Academy of Sciences (Hungary)

Table of Contents

Learning Paraphrasing for Multiword Expressions
Seid Muhie Yimam, Héctor Martínez Alonso, Martin Riedl and Chris Biemann 1

Exploring Long-Term Temporal Trends in the Use of Multiword Expressions
Tal Daniel and Mark Last 11

Lexical Variability and Compositionality: Investigating Idiomaticity with Distributional Semantic Models
Marco Silvio Giuseppe Senaldi, Gianluca E. Lebani and Alessandro Lenci 21

Filtering and Measuring the Intrinsic Quality of Human Compositionality Judgments
Carlos Ramisch, Silvio Cordeiro and Aline Villavicencio 32

Graph-based Clustering of Synonym Senses for German Particle Verbs
Moritz Wittmann, Marion Weller-Di Marco and Sabine Schulte im Walde 38

Accounting ngrams and multi-word terms can improve topic models
Michael Nokel and Natalia Loukachevitch 44

Top a Splitter: Using Distributional Semantics for Improving Compound Splitting
Patrick Ziering, Stefan Müller and Lonneke van der Plas 50

Using Word Embeddings for Improving Statistical Machine Translation of Phrasal Verbs
Kostadin Cholakov and Valia Kordoni 56

Modeling the Non-Substitutability of Multiword Expressions with Distributional Semantics and a Log-Linear Model
Meghdad Farahmand and James Henderson 61

Phrase Representations for Multiword Expressions
Joël Legrand and Ronan Collobert 67

Representing Support Verbs in FrameNet
Miriam R L Petruck and Michael Ellsworth 72

Inherently Pronominal Verbs in Czech: Description and Conversion Based on Treebank Annotation
Zdenka Uresova, Eduard Bejček and Jan Hajic 78

Using collocational features to improve automated scoring of EFL texts
Yves Bestgen 84

A study on the production of collocations by European Portuguese learners
Angela Costa, Luísa Coheur and Teresa Lino 91

Extraction and Recognition of Polish Multiword Expressions using Wikipedia and Finite-State Automata
Paweł Chrząszcz 96

Impact of MWE Resources on Multiword Recognition
Martin Riedl and Chris Biemann 107

A Word Embedding Approach to Identifying Verb-Noun Idiomatic Combinations
Waseem Gharbieh, Virendra Bhavsar and Paul Cook 112

Conference Program

Thursday, 11 August 2016

08:50–09:00 ***Opening remarks***

Oral Session 1

09:00–09:30 *Learning Paraphrasing for Multiword Expressions*
Seid Muhie Yimam, Héctor Martínez Alonso, Martin Riedl and Chris Biemann

09:30–10:00 *Exploring Long-Term Temporal Trends in the Use of Multiword Expressions*
Tal Daniel and Mark Last

10:00–10:30 *Lexical Variability and Compositionality: Investigating Idiomaticity with Distributional Semantic Models*
Marco Silvio Giuseppe Senaldi, Gianluca E. Lebani and Alessandro Lenci

10:30–11:00 ***Coffee Break***

Oral Session 2

11:00–11:20 *Filtering and Measuring the Intrinsic Quality of Human Compositionality Judgments*
Carlos Ramisch, Silvio Cordeiro and Aline Villavicencio

11:20–11:40 *Graph-based Clustering of Synonym Senses for German Particle Verbs*
Moritz Wittmann, Marion Weller-Di Marco and Sabine Schulte im Walde

11:40–12:00 *Accounting ngrams and multi-word terms can improve topic models*
Michael Nokel and Natalia Loukachevitch

12:00–13:00 ***Invited Talk***

13:00–14:00 ***Lunch***

Thursday, 11 August 2016 (continued)

14:00–14:40 Poster Booster Session (5 minutes per poster)

Top a Splitter: Using Distributional Semantics for Improving Compound Splitting
Patrick Ziering, Stefan Müller and Lonneke van der Plas

Using Word Embeddings for Improving Statistical Machine Translation of Phrasal Verbs
Kostadin Cholakov and Valia Kordoni

Modeling the Non-Substitutability of Multiword Expressions with Distributional Semantics and a Log-Linear Model
Meghdad Farahmand and James Henderson

Phrase Representations for Multiword Expressions
Joël Legrand and Ronan Collobert

Representing Support Verbs in FrameNet
Miriam R L Petruck and Michael Ellsworth

Inherently Pronominal Verbs in Czech: Description and Conversion Based on Treebank Annotation
Zdenka Uresova, Eduard Bejček and Jan Hajic

Using collocational features to improve automated scoring of EFL texts
Yves Bestgen

A study on the production of collocations by European Portuguese learners
Angela Costa, Luísa Coheur and Teresa Lino

Thursday, 11 August 2016 (continued)

14:40–15:30 **Poster Session**

15:30–16:00 ***Coffee Break***

Oral Session 3

16:00–16:30 *Extraction and Recognition of Polish Multiword Expressions using Wikipedia and Finite-State Automata*
Paweł Chrząszcz

16:30–16:50 *Impact of MWE Resources on Multiword Recognition*
Martin Riedl and Chris Biemann

16:50–17:10 *A Word Embedding Approach to Identifying Verb-Noun Idiomatic Combinations*
Waseem Gharbieh, Virendra Bhavsar and Paul Cook

17:10–17:20 ***Closing Remarks***