GEMS 2011

GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics

Proceedings of the Workshop

July 31, 2011

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 978-1-937284-16-9 / 1-937284-16-6

Introduction

GEMS 2011 — GEometrical Models of Natural Language Semantics — is the third instalment in a successful series of workshops on distributional models of meaning. Since their earliest application in information retrieval, these models have become omnipresent in contemporary computational linguistics and neighboring fields. Different types of distributional models have been introduced — from the relatively simple bag-of-word, document-based and syntax-based techniques to the statistically more advanced topic models. In the field of lexical semantics, their direct applications include the construction of lexical taxonomies, the recognition of textual entailment, word sense discrimination and disambiguation, cognitive modeling, etc. Moreover, other areas of NLP, like parsing and Machine Translation, have found they can indirectly benefit from the ability of distributional models to generalize from a limited training set to unseen, but semantically similar, words.

The growth of distributional semantics, however, is not without its problems. The aim of GEMS is to address two orthogonal types of current challenges. First, there is the fragmentation with regard to data sets, methods and evaluation metrics, which makes it difficult to compare studies and achieve scientific progress. We addressed this problem by providing authors with two datasets suitable for the evaluation of distributional models, together with the corpora that can be used for their construction. As a result, the performance of very different approaches can be easily compared across papers. Second, these datasets were chosen so as to reflect two of the most pressing issues in the development of distributional models nowadays: differentiation between semantic relations and compositionality.

The first set, presented by Baroni and Lenci, includes concrete nouns from different semantic classes (living, non-living, etc.) with associated words for specific semantic relations such as "attribute", "category coordinate", "event", or "metonym". Panchenko uses this data to compare 21 measures of semantic similarity and relatedness, based on information from WordNet, a traditional corpus, and the web. Baroni, Bruni and Binh Tran explore images as a fourth type of information. Both papers discover fundamental differences in the semantic information that is captured by these different sources of information. This paves the way for a combined, more comprehensive model.

The second dataset, borrowed from Mitchell and Lapata, contains phrase similarity judgments. It makes it possible to address the evaluation of distributional models in compositional tasks. Grefenstette and Sadrzadeh show how a transitive verb can be modeled as a matrix and combine with the vectors of its subject and object. Basile, Caputo and Semeraro use vector permutation in a Random Indexing framework to encode different syntactic dependency relations. Hartung and Frank look to Latent Dirichlet Allocation to identify the dimensions of meaning modified by adjectives.

In addition, the workshop was open to any other original application of distributional semantics. Chan, Callison-Burch and Van Durme use distributional similarity to evaluate paraphrases extracted from a bilingual lexicon. Gulordava and Baroni investigate meaning change in the Google Books corpus.

Obviously, the success of a workshop does not only rely on the quality of its papers. In addition to all speakers and participants, we would like to thank the members of the organizing committee and program committee, who where indispensable for the preparation of the workshop. We also thank our panelists and invited speaker for their thought-provoking contributions, and the ACL SIGSEM and ACL SIGLEX interest groups for their endorsement of the workshop.

Chairs:

Sebastian Padó, University of Heidelberg
Yves Peirsman, University of Leuven & Stanford University

Organizing Committee:

Marco Baroni, University of Trento
Alessandro Lenci, University of Pisa

Program Committee:

Enrique Alfonseca, Google Research
Roberto Basili, University of Roma Tor Vergata
Michael W. Berry, University of Tennessee
Johan Bos, University of Roma La Sapienza
Paul Buitelaar, National University of Ireland
John A. Bullinaria, University of Birmingham
Stefan Evert, University of Osnabrueck
Gregory Grefenstette, Exalead S.A.
Emiliano Guevara, University of Bologna
Matthias Hartung, University of Heidelberg
Iris Hendrickx, University of Lisbon
Kris Heylen, University of Leuven
Jussi Karlgren, Swedish Institute of Computer Science
Zornitsa Kozareva, University of Southern California
Will Lowe, University of Mannheim
Diana McCarthy, Lexical Computing Ltd.
Lukas Michelbacher, University of Stuttgart
Saif Mohammad, University of Maryland
Alessandro Moschitti, University of Trento
Diarmuid O Seaghdha, Cambridge University
Ted Pedersen, University of Minnesota Duluth
Marco Pennacchiotti, Yahoo! Inc.
Daniel Ramage, Stanford University
Lorenza Romano, Fondazione Bruno Kessler
Magnus Sahlgren, Swedish Institute of Computer Science
Sabine Schulte im Walde, University of Stuttgart
Stefan Thater, Saarland University
Peter D. Turney, National Research Council of Canada & University of Ottawa
Tim Van de Cruys, Cambridge University
Yorick Wilks, University of Sheffield
Fabio Massimo Zanzotto, University of Roma Tor Vergata

Invited Speaker:

Mirella Lapata, University of Edinburgh

Panelists:

Roberto Basili & Danilo Croce, University of Roma Tor Vergata
Mona Diab, Columbia University
Annette Frank, University of Heidelberg
Alessandro Lenci, University of Pisa
Peter Turney, National Research Council of Canada & University of Ottawa

Table of Contents

Workshop Program

Sunday, July 31, 2011

8:50–9:00 Opening remarks

9:00–10:00 Invited talk by Mirella Lapata: "Distributional Models of the Representation and Acquisition of Natural Language Categories"

10:00–10:30 *How we BLESSed distributional semantic evaluation*
Marco Baroni and Alessandro Lenci

10:30–11:00 Coffee break

11:00–11:30 *Comparison of the Baseline Knowledge-, Corpus-, and Web-based Similarity Measures for Semantic Relations Extraction*
Alexander Panchenko

11:30–12:00 *Distributional semantics from text and images*
Elia Bruni, Giang Binh Tran and Marco Baroni

12:00–12:30 *Reranking Bilingually Extracted Paraphrases Using Monolingual Distributional Similarity*
Tsz Ping Chan, Chris Callison-Burch and Benjamin Van Durme

12:30–14:00 Lunch

14:00–14:30 *Encoding syntactic dependencies by vector permutation*
Pierpaolo Basile, Annalina Caputo and Giovanni Semeraro

14:30–15:00 *Assessing Interpretable, Attribute-related Meaning Representations for Adjective-Noun Phrases in a Similarity Prediction Task*
Matthias Hartung and Anette Frank

15:00–15:20 *Experimenting with transitive verbs in a DisCoCat*
Edward Grefenstette and Mehrnoosh Sadrzadeh

15:20–15:40 *A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus.*
Kristina Gulordava and Marco Baroni

15:40–16:10 Coffee break

Sunday, July 31, 2011 (continued)

16:10–17:30 Closing session
Invited talk by Peter Turney: "GEMS as the Missing Link between Computational Linguistics and Cognitive Linguistics"
Invited talk by Roberto Basili and Danilo Croce: "Distributional Information, Syntactic Kernels and Compositionality"
Panel discussion