

NAACL HLT 2019

**The 3rd Workshop on Evaluating Vector Space
Representations for NLP**

Proceedings of the Workshop

June 6, 2019
Minneapolis, USA

©2019 The Association for Computational Linguistics

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 978-1-950737-05-5

Preface

The RepEval series of workshops started in the midst of a boom of word embeddings with the goals of promoting new benchmarks for vector space meaning representations, highlighting the issues with existing benchmarks and improving on them. In addition to proposals for new evaluation tasks, it has played an important role by providing an outlet for critical analysis, negative results, and methodological caveats (reproducibility, parameters impact, the issue of attribution of results to the representation or the whole system, dataset structure/balance/representativeness).

Three years later, mainstream NLP is switching to contextualized representations, but we are still facing many of the same issues: reliable intrinsic metrics are scarce, which means that we rarely know what features of representations make them successful for a given downstream task. This makes development of new meaning representations and their fine-tuning a slow and expensive process with too many variables - even more so than before.

The 3rd edition of RepEval aims to foster the discussion of the following issues:

- approaches to intrinsic and extrinsic evaluation of all kinds of distributional meaning representations;
- evaluation motivated by linguistic, psycholinguistic or neurological evidence, its predictive power, and interpretability of meaning representations;
- the (in)stability of vector representations, best practices for reproducible and reliable experiments;
- evaluation of representations at subword level, especially for morphologically complex languages;
- evaluation of phrase, sentence, paragraph and document-level representations: evidence of compositionality, further diagnostic tests, and how much the preservation of abstract syntactic information actually contributes to performance;
- formal analysis of properties of embedding spaces and their impact on downstream tasks;
- the contribution of representations per se vs. other modeling choices to system performance in extrinsic evaluations;
- validation of evaluation methodology and findings in cross-lingual studies;
- specialized vs general-purpose representations, and whether the latter have inherent limits in downstream tasks;
- internal states of end-to-end systems as meaning representations, and ways to make more sense of them.

In the long run, the methodological and practical contributions of RepEval will add to the discussions on what kinds of representations work best for what tasks, how we can interpret and reliably optimize them, and to what extent it is possible to create cross-task meaning representations that would be necessary for general AI.

The third edition of RepEval received 25 submissions (2 more were withdrawn). 4 submissions presented an analysis of existing proposals, 3 contributed proposals for new evaluation tasks, and 3 dealt with scaling, improving or extending prior proposals to other languages. 8 proposals focused on interpretation/analysis of meaning representations, and 7 - on their applications. We accepted 13 submissions, with acceptance rate of 52%.

Organizers:

Anna Rogers, University of Massachusetts Lowell (USA)
Aleksandr Drozd, RIKEN (Japan)
Anna Rumshisky, University of Massachusetts Lowell (USA)
Yoav Goldberg, Bar-Ilan University (Israel)

Program Committee:

Omri Abend, The Hebrew University of Jerusalem (Israel)
Emily Bender, University of Washington (USA)
Sam Bowman, New York University (USA)
Jose Camacho Collados, Cardiff University (UK)
Alexis Conneau, Facebook AI Research (USA)
Barry Devereux, Queen's University Belfast (UK)
Georgiana Dinu, Amazon AWS (USA)
Allyson Ettinger, University of Chicago (USA)
Mohit Iyyer, University of Massachusetts Amherst (USA)
Hila Gonen, Bar-Ilan University (Israel)
Douwe Kiela, Facebook AI Research (USA)
Jonathan K. Kummerfeld, University of Michigan (USA)
Tal Linzen, Johns Hopkins University (USA)
Preslav Nakov, Qatar Computing Research Institute (Qatar)
Neha Nayak, University of Massachusetts Amherst (USA)
Mark Neumann, Allen Institute for Artificial Intelligence (USA)
Denis Paperno, Utrecht University (the Netherlands)
Ellie Pavlick, Brown University (USA)
Marek Rei, University of Cambridge (UK)
Roi Reichart, Technion (Israel)
Vered Shwartz, Bar-Ilan University (Israel)
Diarmuid O'Seaghdha, Apple (UK)
Gabriel Stanovsky, University of Washington (USA)
Karl Stratos, Toyota Technological Institute at Chicago (USA)
Yulia Tsvetkov, Carnegie Mellon University (USA)
Ivan Vulić, University of Cambridge (UK)
Luke Zettlemoyer, University of Washington (USA)

Invited Speakers:

Tal Linzen, Johns Hopkins University (USA)
Kristina Toutanova, Google AI (USA)

Panelists:

Sam Bowman, New York University (USA)
Ryan Cotterell, University of Cambridge (UK)
Barry Devereux, Queen's University Belfast (UK)
Allyson Ettinger, University of Chicago (USA)
Tal Linzen, Johns Hopkins University (USA)

Table of Contents

<i>Neural Vector Conceptualization for Word Vector Space Interpretation</i> Robert Schwarzenberg, Lisa Raithel and David Harbecke	1
<i>Characterizing the Impact of Geometric Properties of Word Embeddings on Task Performance</i> Brendan Whitaker, Denis Newman-Griffis, Aparajita Haldar, Hakan Ferhatosmanoglu and Eric Fosler-Lussier	8
<i>The Influence of Down-Sampling Strategies on SVD Word Embedding Stability</i> Johannes Hellrich, Bernd Kampe and Udo Hahn	18
<i>How Well Do Embedding Models Capture Non-compositionality? A View from Multiword Expressions</i> Navnita Nandakumar, Timothy Baldwin and Bahar Salehi	27
<i>Measuring Semantic Abstraction of Multilingual NMT with Paraphrase Recognition and Generation Tasks</i> Jörg Tiedemann and Yves Scherrer	35
<i>SWOW-8500: Word Association task for Intrinsic Evaluation of Word Embeddings</i> Avijit Thawani, Biplav Srivastava and Anil Singh	43
<i>Classification of Semantic Paraphrasias: Optimization of a Word Embedding Model</i> Katy McKinney-Bock and Steven Bedrick	52
<i>CODAH: An Adversarially-Authored Question Answering Dataset for Common Sense</i> Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez and Doug Downey	63
<i>Syntactic Interchangeability in Word Embedding Models</i> Daniel Hershcovich, Assaf Toledo, Alon Halfon and Noam Slonim	70
<i>Evaluation of Morphological Embeddings for English and Russian Languages</i> Vitaly Romanov and Albina Khusainova	77
<i>Probing Biomedical Embeddings from Language Models</i> Qiao Jin, Bhuwan Dhingra, William Cohen and Xinghua Lu	82
<i>Dyr Bul Shchyl. Proxying Sound Symbolism With Word Embeddings</i> Ivan Yamshchikov, Viascheslav Shibaev and Alexey Tikhonov	90
<i>Multi-Context Term Embeddings: the Use Case of Corpus-based Term Set Expansion</i> Jonathan Mamou, Oren Pereg, Moshe Wasserblat and Ido Dagan	95

Conference Program

Thursday, June 6, 2019

9:00–9:30 *Opening Remarks. Evaluation of meaning representations for NLP: directions and milestones.*

9:30–10:30 *Invited talk: Tal Linzen (Johns Hopkins University)*

10:30–11:00 *Coffee Break*

11:00–12:00 *Invited talk: Kristina Toutanova (Google AI)*

12:00–13:30 *Lunch*

13:30–14:45 *Oral session*

13:30–13:45 *Neural Vector Conceptualization for Word Vector Space Interpretation*
Robert Schwarzenberg, Lisa Raithel and David Harbecke

13:45–14:00 *Characterizing the Impact of Geometric Properties of Word Embeddings on Task Performance*
Brendan Whitaker, Denis Newman-Griffis, Aparajita Haldar, Hakan Ferhatosmanoglu and Eric Fosler-Lussier

14:00–14:15 *The Influence of Down-Sampling Strategies on SVD Word Embedding Stability*
Johannes Hellrich, Bernd Kampe and Udo Hahn

14:15–14:30 *How Well Do Embedding Models Capture Non-compositionality? A View from Multiword Expressions*
Navnita Nandakumar, Timothy Baldwin and Bahar Salehi

14:30–14:45 *Measuring Semantic Abstraction of Multilingual NMT with Paraphrase Recognition and Generation Tasks*
Jörg Tiedemann and Yves Scherrer

14:45–15:00 *1-minute poster madness*

Thursday, June 6, 2019 (continued)

15:00–15:45 *Poster Session*

SWOW-8500: Word Association task for Intrinsic Evaluation of Word Embeddings

Avijit Thawani, Biplav Srivastava and Anil Singh

Classification of Semantic Paraphrasias: Optimization of a Word Embedding Model

Katy McKinney-Bock and Steven Bedrick

CODAH: An Adversarially-Authored Question Answering Dataset for Common Sense

Michael Chen, Mike D’Arcy, Alisa Liu, Jared Fernandez and Doug Downey

Syntactic Interchangeability in Word Embedding Models

Daniel Hershcovich, Assaf Toledo, Alon Halfon and Noam Slonim

Evaluation of Morphological Embeddings for English and Russian Languages

Vitaly Romanov and Albina Khusainova

Probing Biomedical Embeddings from Language Models

Qiao Jin, Bhuwan Dhingra, William Cohen and Xinghua Lu

Dyr Bul Shchyl. Proxying Sound Symbolism With Word Embeddings

Ivan Yamshchikov, Viascheslav Shibaev and Alexey Tikhonov

Multi-Context Term Embeddings: the Use Case of Corpus-based Term Set Expansion

Jonathan Mamou, Oren Pereg, Moshe Wasserblat and Ido Dagan

15:45–16:00 *Coffee break*

16:00–17:15 *A linguist, an NLP engineer, and a psycholinguist walk into a bar... Panel discussion with Sam Bowman, Ryan Cotterell, Barry Devereux, Allyson Ettinger, and Tal Linzen.*

17:15–17:30 *Closing remarks*