ACL 2016

The 54th Annual Meeting of the Association for Computational Linguistics

Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP

August 7-12, 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 978-1-945626-14-2

Introduction

This workshop deals with evaluating vector representations of linguistic units (morphemes, words, phrases, sentences, documents, etc). What marks out these representations - which are colloquially referred to as embeddings – is that they are not trained with a specific application in mind, but rather to capture a characteristic of the data itself. Another way to view their usage is through the lense of transfer learning; the embeddings are trained with one objective, but applied to assist some others. We therefore do not discuss internal representations of deep models that are induced by and applied in the same task.

The Problem with Current Evaluation Methods

Since embeddings are trained in a generally unsupervised setting, it is often difficult to predict their usefulness for a particular task a priori. The best way to assess an embedding's utility is, of course, to use it in a "downstream" application. However, this knowledge tends not to transfer well among different tasks; for example, a 12

To avoid these issues, many papers have chosen to concentrate their evaluation on "intrinsic" (perhaps the more appropriate word is "simple") tasks such as lexical similarity (see, for example: Baroni et al., 2014; Faruqui et al., 2014; Hill et al., 2015; Levy et al., 2015). However, recent work (Schnabel et al., 2015; Tsvetkov et al., 2015) has shown that, just like sophisticated downstream applications, these intrinsic tasks are not accurate predictors of an embedding's utility in other tasks.

One notable issue with current evaluation options is their lack of diversity; despite the large number of intrinsic benchmarks (23 by some counts), and their many differences in size, quality, and domain, the majority of them focus on replicating human ratings of the similarity or relatedness of two words. Even the challenge of analogy recovery through vector arithmetic, which seemed like a more nuanced metric (Mikolov et al., 2013), has been shown to be reducible to a linear combination of lexical similarities (Levy and Goldberg, 2014). As a result, many other interesting linguistic phenomena that are inherent in downstream applications have not received enough attention from the representation learning community.

Goals

New Benchmarks This workshop aims to promote new benchmarks or improvements to existing evaluations that together can address the issues with the existing collection of benchmarks (e.g. lack of diversity). Such benchmarks should fulfill the following criteria:

1. Be simple to code and easy to run

2. Isolate the impact of one representation versus another

3. Improvement in a benchmark should indicate improvement in a downstream application

Better Evaluation Practices The new benchmarks enabled by the workshop will lead to a well-defined set of high quality evaluation resources, covering a diverse range of linguistic/semantic properties that are desirable in representation spaces. Results on these benchmarks will be more easily understood and interpreted by users and reviewers.

Better Embeddings In the long run, the new tasks presented, promoted, and inspired by this workshop should act as a catalyst for faster both technological and scientific progress in representation learning and natural language understanding in general. Specifically, they will drive the development of techniques for learning embeddings that add significant value to downstream applications, and, at the same time, enable a better understanding of the information that they capture.

Submissions

We received 39 submissions, of which 26 were accepted.

Organizers:

Omer Levy, Bar-Ilan University
Felix Hill, University of Cambridge
Anna Korhonen, University of Cambridge
Kyunghyun Cho, New York University
Roi Reichart, Technion - Israel Institute of Technology
Yoav Goldberg, Bar-Ilan University
Antoine Bordes, Facebook AI Research

Program Committee:

Angeliki Lazaridou, University of Trento
Ivan Vulic, Cambridge University
Douwe Kiela, Cambridge University
Torsten Zesch, University of Duisburg-Essen
Preslav Nakov, Qatar Computing Research Institute
Peter Turney, Allen Institute for Artificial Intelligence
German Kruszewski, University of Trento
Manaal Faruqui, Carnegie Mellon University
Karl Stratos, Columbia University
Oren Melamud, Bar-llan University
Minh-Thang Luong, Stanford University
Yulia Tsvetkov, Carnegie Mellon University
Tamara Polajnar, Cambridge University
Laura Rimell, Cambridge University
Marek Rei, Cambridge University
Roy Schwartz, Hebrew University of Jerusalem
Georgiana Dinu, IBM
Omri Abend, Hebrew University of Jerusalem
Antoine Bordes, Facebook AI Research
Mohit Bansal, Toyota Technological Institute at Chicago
Diarmuid O Seaghdha, Vocal IQ
David Jurgens, Stanford University
Alona Fyshe, University of Victoria
Mohit Iyyer, University of Maryland, College Park
Sam Bowman, Stanford University
Neha Nayak, Stanford University
Ellie Pavlick, University of Pennsylvania
Gabriel Stanovsky, Bar-Ilan University

Table of Contents

Conference Program

Friday, August 12th

09:00–09:15 ***Opening Remarks***

09:15–10:00 **Analysis Track**

Intrinsic Evaluation of Word Vectors Fails to Predict Extrinsic Performance
Billy Chiu, Anna Korhonen and Sampo Pyysalo

A critique of word similarity as a method for evaluating distributional semantic models
Miroslav Batchkarov, Thomas Kober, Jeremy Reffin, Julie Weeds and David Weir

Issues in evaluating semantic spaces using word analogies
Tal Linzen

10:00–10:20 **Proposal Track 1**

Evaluating Word Embeddings Using a Representative Suite of Practical Tasks
Neha Nayak, Gabor Angeli and Christopher D. Manning

Story Cloze Evaluator: Vector Space Representation Evaluation by Predicting What Happens Next
Nasrin Mostafazadeh, Lucy Vanderwende, Wen-tau Yih, Pushmeet Kohli and James Allen

10:20–10:45 ***Coffee Break***

Friday, August 12th (continued)

10:45–12:30 **Poster Session**

10:45–11:00 ***Lightning Talks***

Problems With Evaluation of Word Embeddings Using Word Similarity Tasks
Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi and Chris Dyer

Intrinsic Evaluations of Word Embeddings: What Can We Do Better?
Anna Gladkova and Aleksandr Drozd

Find the word that does not belong: A Framework for an Intrinsic Evaluation of Word Vector Representations
José Camacho-Collados and Roberto Navigli

Capturing Discriminative Attributes in a Distributional Space: Task Proposal
Alicia Krebs and Denis Paperno

An Improved Crowdsourcing Based Evaluation Technique for Word Embedding Methods
Farhana Ferdousi Liza and Marek Grzes

Evaluation of acoustic word embeddings
Sahar Ghannay, Yannick Estève, Nathalie Camelin and Paul Deleglise

Evaluating Embeddings using Syntax-based Classification Tasks as a Proxy for Parser Performance
Arne Köhn

Evaluating vector space models using human semantic priming results
Allyson Ettinger and Tal Linzen

Evaluating embeddings on dictionary-based similarity
Judit Ács and Andras Kornai

Evaluating multi-sense embeddings for semantic resolution monolingually and in word translation
Gábor Borbély, Márton Makrai, Dávid Márk Nemeskey and Andras Kornai

Subsumption Preservation as a Comparative Measure for Evaluating Sense-Directed Embeddings
Ali Seyed

Friday, August 12th (continued)

Evaluating Informal-Domain Word Representations With UrbanDictionary
Naomi Saphra

Thematic fit evaluation: an aspect of selectional preferences
Asad Sayeed, Clayton Greenberg and Vera Demberg

12:30–14:00 ***Lunch Break***

14:00–15:30 **Proposal Track 2: Word Representations**

Improving Reliability of Word Similarity Evaluation by Redesigning Annotation Task and Performance Measure
Oded Avraham and Yoav Goldberg

Correlation-based Intrinsic Evaluation of Word Vector Representations
Yulia Tsvetkov, Manaal Faruqui and Chris Dyer

Evaluating word embeddings with fMRI and eye-tracking
Anders Søgaard

Defining Words with Words: Beyond the Distributional Hypothesis
Iuliana-Elena Parasca, Andreas Lukas Rauter, Jack Roper, Aleksandar Rusinov, Guillaume Bouchard, Sebastian Riedel and Pontus Stenetorp

15:30–16:00 ***Coffee Break***

Friday, August 12th (continued)

16:00–17:30 Proposal Track 3: Contextualized Representations

A Proposal for Linguistic Similarity Datasets Based on Commonality Lists
Dmitrijs Milajevs and Sascha Griffiths

Probing for semantic evidence of composition by means of simple classification tasks
Allyson Ettinger, Ahmed Elgohary and Philip Resnik

SLEDDED: A Proposed Dataset of Event Descriptions for Evaluating Phrase Representations
Laura Rimell and Eva Maria Vecchi

Sentence Embedding Evaluation Using Pyramid Annotation
Tal Baumel, Raphael Cohen and Michael Elhadad

17:30–18:15 ***Open Discussion***

18:15–18:30 ***Best Proposal Awards***