Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)
Doctoral Consortium Track

A Unified Framework for Intrinsic Evaluation of Word-Embedding Algorithms

Anna Giabelli
Università degli studi di Milano - Bicocca, Italy
anna.giabelli@unimib.it

Abstract

Word embeddings are widely used in copious Natural Language Processing tasks, including semantic analysis, information retrieval, dependency parsing, question answering, and machine translation. This extensive use implies that the evaluation of the performance of such representations is crucial for choosing the best model to perform those tasks.

Though there are well-established procedures and benchmarks for intrinsic evaluation, as far as we know, a unified method of evaluation that can merge the results of those tasks to provide a comprehensive evaluation is missing. The main goal of this work is to create a pipeline to blend all major intrinsic evaluation tasks to compute such overall evaluation - the *PCE* - of word embeddings.

1 Introduction

Word embeddings are vector representations of words based on the hypothesis that words occurring in a similar context are prone to have a similar meaning. Words are represented by semantic vectors, which are usually derived from a large corpus using co-occurrence statistics, and their use improves learning algorithms in many NLP tasks. Learning high-quality representations of words is extremely important for these tasks, yet the question of what can be considered a good word embedding model remains an open problem.

In [Schnabel *et al.*, 2015] the authors claim that a good embedding provides vector representations of words such that the relationship between two vectors mirrors the linguistic relation between the two terms. As a consequence, a method of word embeddings evaluation is any technique of finding embeddings correlation with any data that hypothetically could carry information about lexical semantics [Bakarov, 2018].

The existing approaches can be divided into two major classes: intrinsic and extrinsic. *Intrinsic evaluations* of word embeddings directly test for *syntactic or semantic relationships* between words [Baroni *et al.*, 2014]. These tasks typically involve a pre-selected set of query terms, and semantically related target words, with human judgments on word relations [Schnabel *et al.*, 2015]. Methods of *extrinsic evaluation* are based on the ability of word embeddings to be used as the feature vectors of supervised machine learning algorithms; the performance of the supervised model acts as a measure of word embeddings quality.

1.1 Motivation

One of the most notable questions in distributional semantics is how to evaluate the quality of word embeddings. Still, there is no consensus about what an *effective* assessment measure is, and furthermore, there are no perfect evaluation methods because it is difficult to understand how the embedding spaces encode linguistic relations [Wang *et al.*, 2019].

Since word embeddings are widely used in copious Natural Language Processing tasks, and given the high sensitivity of word embeddings to small changes in hyper-parameters during their generation [Levy *et al.*, 2015], it is crucial to have a reliable measure of evaluation, but now there are just different tasks that evaluate different aspects of the models, and there is not a method for merging those evaluations into a comprehensive one. My goal is to create this overall evaluation metric - the *PCE* - for word embedding models.

2 State of the Art

There is a large amount of literature on the existing intrinsic evaluation methods for word embeddings models. In [Schnabel *et al.*, 2015] the authors conduct the first comprehensive study covering a wide range of evaluation criteria and popular embedding techniques. They perform a comprehensive analysis of evaluation methods and introduce novel ones, providing insights into the strengths of different embeddings.

Probably the most used metric for intrinsic evaluation is *Semantic Relatedness* [Baroni *et al.*, 2014], which evaluates the performance of a vector model by the correlation between the cosine similarity between pairs of word vectors and a human scores of relatedness between them. A similar method is *Semantic Similarity*, which reduces relatedness to similarity.

Another well-known task is *Analogy*, which was popularized by [Mikolov *et al.*, 2013]: the goal is to find a term x for a word y so that $x : y$ best resembles a relationship $a : b$.

The third most used task is *Categorization*: given a set of nominal concepts, the objective is to group them into natural categories (e.g., dogs and cats into the mammal class) [Baroni *et al.*, 2014]. The word vectors are clustered into groups, and the performance is usually evaluated in terms of purity.

There are several benchmarks available for these methods.

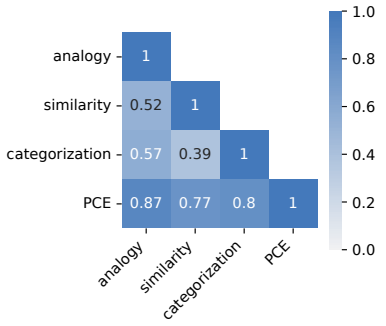

Figure 1: Correlation between *PCE* and the intrinsic tasks.

3 Objectives of the Research

The goal of my work is to create a method to blend the main intrinsic evaluation methods and to generate a comprehensive evaluation metric of word embedding models - the *PCE* - based on the performance achieved by those models on state-of-the-art benchmark tasks and data.

As for now, I consider three state-of-the-art tasks: word similarity, word analogy, and concept categorization. The choice to initially focus on these three tasks follows from what Wang et al. [2019] found in their study of performance consistency of extrinsic and intrinsic evaluators. They found word similarity, analogy, and concept categorization to be the more effective intrinsic evaluators. Since those tasks can perform differently for different downstream tasks, they suggest using these three evaluators jointly when testing a new word embedding model. Starting from the performance of the word embedding models on the benchmarks available for those three tasks, the overall evaluation for a task is computed as the mean of the evaluation over its benchmarks. *PCE*, the final evaluation, is computed as the first principal component obtained through Principal Component Analysis (PCA).

Given the standardized variables $\mathbf{z}_1, \ldots, \mathbf{z}_p$ and their covariance $\mathbf{R}$, the eigenvectors $\mathbf{a}_k$ of the correlation matrix $\mathbf{R}$ define the $q < p$ uncorrelated maximum-variance linear combinations, that are called the principal components (PCs). By keeping just the first PC - the one that explains most of the variance in the original p variables - we can obtain:

$$PCE = \mathbf{Z}\mathbf{a}_1 = \sum_{j=1}^{p} a_{j1}\mathbf{z}_j \tag{1}$$

The *PCE* is normalized to make it easier to compare the evaluation of the different models. Figure 1 shows that the correlation of the evaluation for the three tasks with the *PCE* is high, ranging from 0.77 to 0.87. The three tasks - analogy, similarity, and categorization - show a less strong correlation with each another, leading us to think that, by using them for computing the *PCE*, we are not considering redundant information, and each task contributes valuable information.

I am considering only intrinsic evaluators because extrinsic evaluation is suited to assess the performance on a specific downstream task, but when the embeddings are used in a wide range of different tasks, it fails to provide a global evaluation since word embeddings' performance scores in various downstream tasks do not correlate [Bakarov, 2018].

At the moment, I am not considering contextual word embeddings, e.g., BERT, for two reasons. The first reason is that state-of-the-art evaluation tasks are less applicable to such embeddings than classic mono-sense embeddings. For example, for evaluating BERT on a benchmark for semantic similarity, one should consider just one a-contextual meaning for the words in the dataset, not taking into account the advantage of contextual word embeddings. The second one is that dynamic embeddings are more computationally expensive to train than static embeddings, requiring a massive amount of data and computational power [Roy and Pan, 2021].

4 Conclusions and Future Contributions

My future work is directed at extending this framework, including further intrinsic evaluation methods, like synonym detection [Baroni *et al.*, 2014], or HSS [Malandri *et al.*, 2020; Giabelli *et al.*, 2022]. I also aim to find a way to include the evaluation of contextual word embeddings, considering tasks specific for their evaluation.

Acknowledgments

This work has been achieved thanks to prof. Fabio Mercorio, prof. Mario Mezzanzanica, and Dr. Lorenzo Malandri.

References

[Bakarov, 2018] A. Bakarov. A survey of word embeddings evaluation methods. *arXiv:1801.09536*, 2018.

[Baroni *et al.*, 2014] M. Baroni, G. Dinu, and G. Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL*, 2014.

[Giabelli *et al.*, 2022] A. Giabelli, L. Malandri, F. Mercorio, M. Mezzanzanica, and N. Nobani. Embeddings evaluation using a novel measure of semantic similarity. *Cognitive Computation*, 2022.

[Levy *et al.*, 2015] O. Levy, Y. Goldberg, and I. Dagan. Improving distributional similarity with lessons learned from word embeddings. *TACL*, 3, 2015.

[Malandri *et al.*, 2020] L. Malandri, F. Mercorio, M. Mezzanzanica, and N. Nobani. Meet: A method for embeddings evaluation for taxonomic data. In *2020 ICDMW*. IEEE, 2020.

[Mikolov *et al.*, 2013] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 2013.

[Roy and Pan, 2021] A. Roy and S. Pan. Incorporating extra knowledge to enhance word embedding. In *IJCAI*, 2021.

[Schnabel *et al.*, 2015] T. Schnabel, I. Labutov, D. Mimno, and T. Joachims. Evaluation methods for unsupervised word embeddings. In *EMNLP*, 2015.

[Wang *et al.*, 2019] B. Wang, A. Wang, F. Chen, Y. Wang, and C. C. J. Kuo. Evaluating word embedding models: Methods and experimental results. *APSIPA*, 8, 2019.