Big Data in Public and Social Services
February 2021 Exam session

Who Evaluates the Evaluator? A Survey of Word Embedding Evaluation Methods

Sanja Stanišić, Università degli Studi di Milano Bicocca, CdLM Data Science, matr. 800409

Abstract
Word embedding is a real-valued vector representation of words able to grasp lexical semantic from a natural language corpus. Word embedding models have been extensively used in natural language processing (NLP) tasks. Since they capture complex language characteristics, their evaluation is very difficult. There is no universal word embedding evaluator, embedding methods should be compared in the context of a specific task. This paper is aimed at summarizing the state-of-the art evaluation methods.

Keywords
Word embedding, word embedding evaluation, natural language processing

Contents

1. Introduction

Word embedding is a real-valued vector representation of words capable of capturing their semantic and syntactic meaning. Embeddings are a valuable tool in solving numerous natural language processing (NLP) tasks, including semantic role labeling, part-of-speech tagging, chunking etc. A good embedding provides vector representations of words such that the relationship between two vectors resembles the linguistic relationship between two words [1].

A word embedding is considered as mapping:

$$\mathrm{V} \rightarrow \mathrm{R}^{\mathrm{D}} : w \rightarrow \vec{w}$$

that maps a word w from a vocabulary V to a real-valued vector $\vec{w}$ in an embedding space of dimensionality D.

Methods to generate this mapping are different, they mainly use distributional semantic models (DSMs). These models are based on the distributional hypothesis which suggests that the more semantically similar two words are, the more distributionally similar they will be in turn, and thus the more they will tend to occur in similar linguistic contexts.

Vectors' similarity is commonly measured by cosine similarity defined as: $similarity\ (w_1\ , w_2\ ) = \frac{\overrightarrow{w_1} * \overrightarrow{w_2}}{\|\overrightarrow{w_1}\| \|\overrightarrow{w_2}\|}$

The list of nearest neighbors of a word w are all words $v \in V \setminus \{w\}$, sorted in descending order by similarity (w, v), and w is considered target word.

Research in computational linguistics has shown that contextual information provides a good approximation to word meaning, since semantically similar words tend to have similar context distributions. DSMs use vectors to keep track of the context (e.g. co-occurring words) in which the target word appears [2]. In the past decade new probabilistic architecture was introduced, such as: Countinuous Bag-of-Words Model – CBOW and Continuous Skip-gram Model [3]. Experimental results of Baroni at al. [2], who systematically compared count and predict models of vector representations, demonstrated that predict models performed better and there was a very good reason to switch to the then new architecture capable of learning high-quality word vectors from huge datasets.

So, different embedding models produce different vector representations. They initially aimed at representing each word type as a single point in the semantic space. This objective created a major limitation: embedding models ignored the fact that words could have multiple meanings and conflated all these meanings into a single representation. Despite their flexibility and success in capturing semantic properties of words, in order to be effective embeddings had to overcome the so-called ***meaning conflation deficiency*** [4]. A solution to this shortcoming was to represent individual meanings of words, i.e. word senses, as independent representations (*sense representations*).

Two different approaches were adopted in learning sense representation. These approaches have different sources of learning word senses and they model them in a different way. ***Unsupervised models*** learn word senses directly from text corpora (this paradigm is very related to sense induction). On the other hand, ***knowledge-based models*** represent word senses as defined by an external sense inventory, i.e. they exploit the sense inventories of lexical resources (the most associated task in this case is word sense disambiguation) [4].

Therefore, having in mind all of the above, there are several properties that ***all*** embedding models should aim for [5]:

- **Non-conflation**

  Embedding models should be able to recognize differences in the contexts around a word and encode these details (e.g. plural/singular form, tenses etc.) into a meaningful representation in the word subspace.

- **Robustness against lexical ambiguity**

  All meanings (or senses) of the word should be represented. Models should be capable of recognizing the meaning of the word from its context and find the appropriate embedding. For example, an embedding should be able to represent the difference between the following: “the **bow** of a ship” and “**bow** and arrows”.

- **Multifacetedness**

  Different phonetic, morphological, syntactic and other properties of a word should contribute to its final representation. For example, the representation of a word should change when the tense is changed or a prefix is added.

- **Reliability**

  Results of a word embedding model should be reliable. This is important as word vectors are randomly initialized when being trained. Even if a model creates different representations from the same dataset because of random initialization, the performance of various representations should score consistently.

- **Good geometry**

  The embedding space geometry should have a good spread. It means that a smaller set of more frequent, unrelated words should be evenly distributed throughout the space, while a larger set of rare words should cluster around frequent words. Models should overcome the difficulties arising from the inconsistent frequency of word usage and derive some meaning from the frequency.

2. Evaluation of Embedding Models

Evaluation of embedding models aims at comparing their characteristics using quantitative representative metric. However, it is not easy to identify a concrete and uniform way of evaluating these abstract characteristics, and therefore defining a good embedding model is still an open problem [5].

Different evaluations result in different orderings of the embedding models, casting doubt on the common assumption that there is one single optimal vector representation [1]. Hence, there is no universal word embedding evaluator.

In 2015 the approaches to evaluation of word embeddings were systematized [1] in two major categories: extrinsic and intrinsic evaluation. **Extrinsic evaluation** uses word embeddings as input features to a downstream NLP task, and subsequently measures changes in performance metrics specific to that task. On the other hand, **intrinsic evaluation** directly tests syntactic or semantic relationship between words.

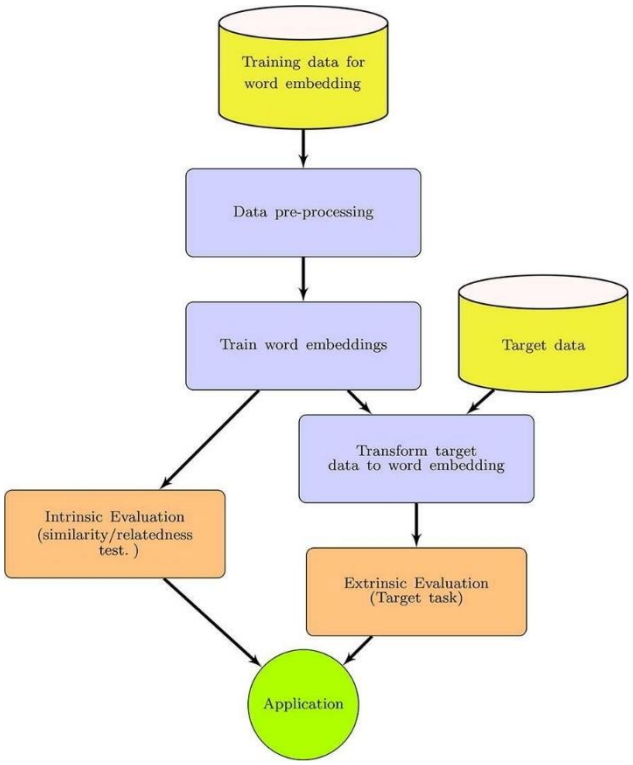

Figure 1 : Extrinsic and Intrinsic Evaluation

Figure 1 [6] illustrate the steps leading to extrinsic and intrinsic evaluation.

More recent research papers suggest an additional evaluation category: **global evaluation** [7]. These global methods are data-free (i.e. with no external data needed other than the word embedding matrix) evaluations investigating relationships between vectors or studying their distribution. Measures proposed in [7] are *global intrinsic dimensionality* and *effective rank and empirical dimension.*

An exhaustive classification of all existing word embedding evaluation methods, both widely used and experimental, was introduced in [8] and will be presented in next chapters.

As stated in [5], a good word embedding evaluator should strive to have the following properties:

- **Good testing data**

  Testing data should be varied with a good spread in the span of a word space, in order to ensure a reliable representative score. Both frequently and rarely occurring words should be included in the evaluation.

- **Comprehensiveness**

  An evaluation method should ideally test for many properties of a word embedding.

- **High correlation**

  The score of an embedding model obtained by intrinsic evaluation should be well correlated with the performance of the model in solving a downstream NLP task.

- **Efficiency**

  Evaluation methods should be computationally efficient. Although most embeddings are created to solve computationally expensive downstream tasks, their evaluators should be simple, yet able to predict the downstream performance of an embedding.

- **Statistical significance**

  The evaluation results of different embedding models by a single evaluator should have enough statistical significance. In other words, there should be enough variance between score distribution in order to have the embeddings differentiated.

3. Extrinsic Evaluation Methods

Extrinsic evaluation methods use word embeddings as feature vectors for supervised machine learning algorithms aimed at solving various downstream NLP tasks. The performance of a supervised model represents the measure of the word embedding quality. Theoretically, any NLP downstream task could be considered as extrinsic evaluation method. Here, only a few methods will be presented:

Part-of-Speech (POS) Tagging aims to assign POS tags to each word in the sentence like noun, verb, adverb,

adjective etc. It is used for several purposes such as text indexing and retrieval.

Chunking (Shallow Parsing), similarly to POS, firstly assigns tag to each word indicating its proprieties (e.g. verb, noun, adjective etc.), and then group as chunks syntactically correlated words (verb phrases, noun phrases etc.). In comparison with POS tagging, chunking provides more information about the structure of the sentence or phrases in the sentence.

Named Entity Recognition (NER) focuses on identifying named entities (e.g. names of persons, locations, organizations etc.) and numeric expressions (e.g. time, percentage etc.). Therefore, it tests the capacity of the model to extract high-level information from a plain text data.

Text Classification is aimed at marking a text fragment with a label which reflects it context. For example, sport news can be categorized with reference to the type of sport they are reporting about.

Sentiment Analysis is a particular text classification problem. In general, a text fragment is marked with a binary/multi-level label representing positive or negative polarity of the text's sentiment. Word phrases play very important role in taking the final decision – a negation such as "no" or "not" can reverse the meaning of the whole sentence.

Extrinsic evaluation makes sense if the comparison of web embedding models in solving one specific downstream task is needed. In that case the evaluation of the performance of the supervised model will give the adequate score of the quality of the embeddings. However, as proven in [1] different tasks favor different embeddings, therefore extrinsic evaluation methods cannot be used as absolute measure of word embeddings quality.

4. Intrinsic Evaluation Methods

Intrinsic Evaluation Methods measure directly syntactic or semantic relationship between words. These methods compare word embeddings with human judgement on word relations. Typically, sets of words created manually are used to obtain human assessments which are then compared to word embeddings [8]. Gathering of human assessments is conducted either directly on a limited number of real users, or through pre-collected offline data (using crowdsourcing platforms with unlimited number of participants). An aggregate score, such as correlation coefficient, is calculated for each of the embeddings and it serve as an **absolute measure of embedding quality**.

In **comparative intrinsic evaluation** the assessors give direct feedback on embeddings themselves – when different embedding models produce different judgement on relationships between words, the assessor has to evaluate which one works better. Therefore, comparative intrinsic evaluation doesn't provide an absolute measure of embedding's quality, but allow the selection of the most adequate embeddings in a given set.

Comparative intrinsic methods of evaluation were introduced in [1]. The authors proposed this new approach to evaluation jointly with the different formulation of tasks presented to annotators in charge of assessing the embedding models: all tasks were formulated as choice problem rather than ordinal relevance task, which made easier the work of the annotators. Experimental results conducted in [1] have shown that there is strong correlation between automated similarity evaluation and direct human evaluation i.e. that data gathered through crowdsorcing are as efficient as on-line collected data, at least for the similarity task.

The same paper [1] has shown how outcomes of different evaluation criteria (word relatedness, coherence, downstream performance) are connected – experimental results have indicated that the embedding methods should be compared in the context of a specific task, e.g. linguistic insight or good downstream performance.

Therefore, intrinsic evaluations, unless they are coupled with an extrinsic task, have little value in themselves. The main purpose of an intrinsic evaluator is to serve as a *proxy* for a downstream task embeddings are tailored for [9].

Yet another attempt in systemizing all existing intrinsic method of evaluation, both widely-used and experimental, was made by Bakarov [8] who proposed four groups of intrinsic evaluations:

1. **Intrinsic conscious evaluation**
   This group encompasses those methods which, inspired by data collection in psycholinguistic research, collect the assessments that are results of a conscious process in a human brain (i.e. assessors have time to think about their

answers). The human judgement might be biased by subjective factors (e.g. when there isn't a clear definition of meaning, every person interprets the relationship between words in its own way, introducing the variability to the estimates). Therefore, it is not clear if the conscious assessments are really able to report the structure of semantics in a natural language.

2. **Intrinsic subconscious evaluation**
   Evaluation methods in which the human assessment is conducted through collection of reflective responses are classified in this group. The process of evaluation here is more complex and interdisciplinary as it is based on using various neuroimaging methods, previously used only in psycholinguistics.
   These methods are based on scientific research in psycholinguistics which indicates that sematic structure of a natural language lies in the so-called subconscious level of cognition [10]. If the assessment could be collected at this level, the results would be far less biased then the assessment resulting from a conscious process.

3. **Intrinsic thesaurus-based evaluation**
   This group of evaluation methods doesn't compare word embedding models to human assessment, but to knowledge bases, semantic networks and manually constructed thesauri.

4. **Intrinsic language driven evaluation**
   This group comprises of methods based on comparison of word embedding models to data underlying in a language itself. Such data is found, for instance, in graphematic representation of words, speech sound signals or occurrence frequency of word pairs in a corpus.

Following are the examples of evaluators for each one of the classification groups.

4.1 Intrinsic Conscious Evaluation

Word similarity

This method is one of the most widely used methods which compares the distance between words in an embedding space and human assessment of the actual semantic distance between words. The more similar are the two distances, the better is the embedding model.

This evaluation method is, at the same time, one of the most popular and most frequently criticized. In particular, it has been criticized for the (un)reliability of human linguistic judgements, which are subject to over 50 potential linguistic, psychological, and social confounds. The ambiguity in defining semantic similarity is also anissue (which often may be confused with relatedness: *car* and *train* are two similar words, while *car* and *road* are two related words), and the rating of relatedness is particularly confusing [11].

Word analogy

This method also goes by name *analogical reasoning*, *linguistic regularities* and *word semantic coherence*. It is based on the idea that arithmetic operations in a word vector space could be predicted by humans: given a pair of words a and a^*, and a third word b , the analogy relationship between a and a^* can be used to find the corresponding b^* such as:

$$a : a^* = b : b^* .$$

For example:

write : writing = read : reading,

where reading is the word to be found by analogy.

There are several metrics used in the word analogy task:

- *3CosAdd* (and similar *3CosMul*) based on arithmetic operations in vector space (addition and multiplication of cosine distance)
- *PairDir* modifies *3CosAdd* taking into account the direction of the resulting vector
- *Analogy Space Evaluation* compares the distances between words directly without finding the nearest neighbors.

Some researches indicated that many model score poorly on the analogy test, suggesting that not all relations can be identified in this way. In particular, the problem is evident with synonyms and antonyms.

Another problem with this test is human subjectivity. Analogies are fundamental to human reasoning and logic. The dataset on which the current models are trained doesn't encode our sense of reasoning, and it is rather different from how humans learn natural languages. Thus, given a word pair, vector space model may find a different relationship from what humans may find [5].

Concept categorization, also called *word clustering*, is somewhat different from the previous two models. It is aimed at splitting a given set of words into subsets of words belonging to different categories – for example, if a

set is composed of words *dog*, *elephant*, *robin*, *crow*, the first two would make the cluster of *mammals*, while the last two would form the cluster of *birds*. The number of clusters is defined.

The test is conducted in a following way: first the vector corresponding to each word is calculated; then a clustering algorithm (such as the *k means* algorithm) is used to separate the set of word vectors into *n* different categories; a performance metric is defined based on the class purity, where purity refers to whether each cluster contains concepts form the same or different categories.

There are several challenges that this model is facing, from not having the standardized splits of datasets to not having specific clustering algorithm defined. However, one major issue is again human subjectivity, as humans can group words using concepts that embeddings would gloss over. Given the words *banana, sun, lemon, blueberry, ocean, iris,* one can group them into yellow objects (*lemon, sun, banana*) and blue objects (*blueberry, iris, ocean*). But, since words can belong to different groups we can have (*lemon, banana, iris, blueberry)* clustered as plants, and (*sun, ocean*) belonging to nature category. We, actually, do not have adequate method to evaluate the quality of each cluster [5]. One good property of this model is its ability to test for the frequency effect, since it is good in revealing where frequent words are clustered together.

Outlier Detection

This is a relatively new method which also evaluates clustering in vector space, only this time the task is not to divide words in a certain amount of clusters, but to find words that do not belong in an already formatted cluster. For example, if the following words are clustered in one group {*orange, banana, lemon, book, orange*} where there are mostly fruits, the word *book* is the outlier, since it is not a fruit.

Therefore, this evaluator tests the semantic coherence of vector space models, where semantic clusters can be first identified. If frequent words are clustered to form hubs, while rarer words are not clustered around the more frequent words they relate to, the evaluator will not perform well in this metric.

This evaluator relies heavily on human reasoning and logic. The outliers identified by humans are strongly influenced by characteristics of words perceived to be important [5].

4.2 Intrinsic Subconscious Evaluation

Semantic priming

This evaluation method is based on the same psycholinguistic experiment. It hypothesizes that a person reads a word faster if it is preceded by a semantically related word. Therefore, the idea of the experiment is to measure the time of reading a target word *a*, when it occurs after a word b_1 and when it occurs after a word b_2. If the reading time is lower when b_1 precedes the target word, then the word b_1 is claimed to be semantically related to *a* (it is called *prime/prime word/stimulus word*). The time of reading could be obtained with the help of eye-tracking or self-paced reading [8].

Neural Activation Pattern

This is yet another experimental method, not in wide use as well as all methods belonging to this group. The underlying idea hypothesizes that word meanings are reflected in some patterns in the brain, and such patterns could be used as input data for embeddings. Therefore, exploration of the neural activation patterns could help in better understanding of the nature of semantics, and therefore to consistently evaluate existing methods of embeddings evaluation.

However, the consistency of such brain data is questioned, since neural activation patterns do not correlate in large number of subjects (because of the different size and structure of their brains).

The proposed evaluation methods could be based on functional magnetic resonance imaging (fMRI) or electroencephalography (EEG) [8].

4.3 Intrinsic Thesaurus-Based Evaluation

QVEC

This evaluation method is based on correlations of learned vectors with linguistic word vectors constructed from rich linguistic resources, annotated by domain experts. The hypothesis is that dimensions in the distributional vectors correspond to linguistic properties of words. Linear combinations of vector dimensions are expected to produce relevant content.

The major issue of this method is the subjectivity of man-made linguistic vectors. Current word embeddings perform much better than man-made models as they are based on statistical relations from data [5].

4.4 Intrinsic Language Driven Evaluation

Phonosemantic analysis

This method is based on the idea that the form of the linguistic sign is not arbitrary, since it somehow correlates with its semantics. If this was true, then is would be possible to obtain data pertinent to semantics of a word trough phonosemantic patterns of its phonemes or characters. In order to calculate phonosemantic difference between two words Levenshtein distance could be measured (a string metric for measuring difference between two sequences) and use it as a gold standard.

5. Conclusion

Word embeddings capture complex language characteristics and their evaluation is difficult. Since there is no such thing as one single optimal vector representation, different evaluations result in different orderings of the embedding models, and therefore there is no universal word embedding evaluator.

Study of Wang at al. [5] offer valuable guidance is selecting suitable evaluation methods for different application tasks as they conducted most complete experimental extrinsic and intrinsic evaluations on six widely used embedding models (**SGNS** – skip gram with negative sampling, **CBOW** – continuous bag-of-words, **GloVe** - global vector for word representation, **FastText**, **ngram2vec** and **Dic2vec**), using benchmark datasets for each task. They also performed a consistency study of extrinsic and intrinsic evaluators using correlation analysis.

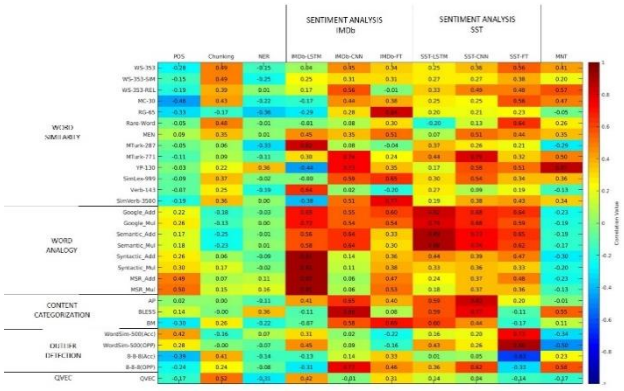

Figure 2: Pearson's correlation between intrinsic and extrinsic evaluators

As an example of their comprehensive experimental results, Figure 2 [5] shows Pearson's correlation between intrinsic and extrinsic evaluation models. Extrinsic evaluators are on x-axis, while intrinsic ones are on y-axis. The shades of colors represent the degree of correlation ranging from darkest blue, which represent the maximum negative correlation, to dark red, representing the maximum positive correlation.

To conclude, there is no such thing as perfect embedding model, because none of the tested is consistent in performing well in all tasks.

6. References

[1] T. Schnabel, I. Labutov, D. Mimno and T. Joachims, "Evaluation methods for unsupervised word embeddings," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, 2015.

[2] M. Baroni, G. Dinu and G. Kruszewski, "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, 2014.

[3] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," 2013.

[4] J. Camacho - Collados and M. T. Pilehvar, "From Word to Sense Embeddings: A Survey on Vector Representations of Meaning," *Journal of Artificial Intelligence Research,* vol. 63, pp. 743-788, 2018.

[5] B. Wang, A. Wang, F. Chen, Y. Wang and C. -. C. J. Kuo, "Evaluating word embedding models: methods and experimental results," *APSIPA Transactions on Signal and Information Processing,* vol. 8, no. E19, 2019.

[6] F. Khan Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney and F. Rudzicz, "A survey of word embeddings for clinical text," *Jpurnal of Biomedical Informatics ,* vol. 100, no. Supplement, 2019.

[7] F. Torregrossa, V. Claveau, N. Kooli, G. Gravier and R. Allesiardo, "On the Correlation of Word Embedding Evaluation Metrics," in *Proceedings of*

the 12th Conference on Language Resources and Evaluation (LREC 2020),, Marseille, 2020.

[8] A. Bakarov, "A Survey of Word Embeddings Evaluation Methods," arXiv, 2018.

[9] Y. Tsvetkov, M. Faruqui and D. Chris, "Correlation-based Intrinsic Evaluation of Word Vector Representations," Berlin, 2016.

[10] M. Kutas and K. D. Federmeier, "Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP)," *Annual review of psychology,* vol. 62, pp. 624-647, 2011.

[11] A. Gladkova and A. Drozd, "Intrinsic Evaluations of Word Embeddings: What Can We Do Better?," in *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*, Berlin, 2016.

[12] S. Ghannay, B. Favre, Y. Estève and N. Camelin, "Word Embeddings Evaluation and Combination," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portoroz, 2016.