Predicting Word Embeddings Variability

Bénédicte Pierrejean and **Ludovic Tanguy**
CLLE: CNRS & University of Toulouse
Toulouse, France
{benedicte.pierrejean,ludovic.tanguy}@univ-tlse2.fr

Abstract

Neural word embeddings models (such as those built with *word2vec*) are known to have stability problems: when retraining a model with the exact same hyperparameters, words neighborhoods may change. We propose a method to estimate such variation, based on the overlap of neighbors of a given word in two models trained with identical hyperparameters. We show that this inherent variation is not negligible, and that it does not affect every word in the same way. We examine the influence of several features that are intrinsic to a word, corpus or embedding model and provide a methodology that can predict the variability (and as such, reliability) of a word representation in a semantic vector space.

1 Introduction

Word embeddings are dense representations of the meaning of words that are efficient and easy to use. Embeddings training methods such as *word2vec* (Mikolov et al., 2013), are based on neural networks methods that imply random processes (initialization of the network, sampling, etc.). As such, they display stability problems (Hellrich and Hahn, 2016) meaning that retraining a model with the exact same hyperparameters will give different word representations, with a word possibly having different nearest neighbors from one model to the other.

Benchmarks test sets such as WordSim-353 (Finkelstein et al., 2002) are commonly used to evaluate word embeddings since they provide a fast and easy way to quickly evaluate a model (Nayak et al., 2016). However, the instability of word embeddings is not detected by these test sets since only selected pairs of words are evaluated. A model showing instability could get very similar performance results when evaluated on such benchmarks.

Hyperparameters selected when training word embeddings impact the semantic representation of a word. Among these hyperparameters we find some hyperparameters internal to the system such as the architecture used, the size of the context window or the dimensions of the vectors as well as some external hyperparameters such as the corpus used for training (Asr et al., 2016; Baroni et al., 2014; Chiu et al., 2016; Li et al., 2017; Melamud et al., 2016; Roberts, 2016). In this work, we adopt a corpus linguistics approach in a similar way to Antoniak and Mimno (2018), Hamilton et al. (2016) and Hellrich and Hahn (2016) meaning that observing the semantic representation of a word consists in observing the nearest neighbors of this word. Corpus tools such as Sketch Engine (Kilgarriff et al., 2014) use embeddings trained on several corpora[1] to provide users with most similar words as a lexical semantic information on a target word. In order to make accurate observations, it thus seems important to understand the stability of these embeddings.

In this paper, we measure the variation that exists between several models trained with the same hyperparameters in terms of nearest neighbors for all words in a corpus. A word having the same nearest neighbors across several models is considered stable.

Based on a set of selected features, we also attempt to predict the stability of a word. Such a prediction is interesting to understand what features have an impact on a word representation variability. It could also be used to certify the reliability of the given semantic representation of a word without having to retrain several models to make sure the representation is accurate. This will be a useful method to give more reliability to observations made in corpus linguistics using word em-

[1] https://embeddings.sketchengine.co.uk/static/index.html

*Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 154–159
New Orleans, June 5-6, 2018.

beddings. It can also help choosing the right hyperparameters or refine a model (e.g. by removing selected semantic classes).

We examine the influence of several features that are intrinsic to a word, a corpus or a model: part of speech (henceforth POS), degree of polysemy, frequency of a word, distribution of the contexts of a word, position and environment of a vector in the semantic space. We train a multilinear regression model using these features and predict up to 48% of the variance. This experiment was conducted on 3 different corpora with similar results. We first explain how we measure the variation of a model. We then present the models used in this work and we finally describe our predictive model.

2 Experiment Setup

To measure the variation for a word between two embedding models, we used an approach similar to Sahlgren (2006) by measuring the nearest neighbors overlap for words common to the two models. More precisely the variation score of a word *varnn* across two models M_1 and M_2 is measured as:

$$varnn^{N}_{M_1,M_2}(w) = 1 - \frac{|nn^{N}_{M_1}(w) \cap nn^{N}_{M_2}(w)|}{N}$$

$nn^{N}_{M}(w)$ represents the N words having the closest cosine similarity score with word w in a distributional model M. In the experiments presented here we selected $N = 25$. To chose the value of N, we selected two models and computed the variation with different values of N across the entire vocabulary (1, 5, 10, 25, 50 and 100). We then computed the correlation coefficient between scores for all the N values and found that the highest average correlation value was for $N = 25$. The variation was computed only for open classes (adverbs, adjectives, verbs and nouns).

This variation measure presents both advantages and inconvenients. The fact that this measure is cost-effective and intuitive makes it very convenient to use. It is also strongly related to the way we observe word embeddings in a corpus-linguistics approach (i.e. by observing a few nearest neighbors). However we are aware that this measure assess only a part of what has changed from one model to the other based on the number of neighbors observed. This measure may also be sensible to complex effects and phenomena in high-dimensional vector spaces such as *hubness*, with some words being more "popular" nearest neighbors than others (Radovanović et al., 2010). Although we could indeed identify such hubs in our vector spaces, they were limited to a small cluster of words (such as surnames for the BNC) and did not interfere with our measure of stability for all other areas of the lexicon.

The compared models were trained using the standard *word2vec*[2] with the default hyperparameters (architecture Skip-Gram with negative sampling rate of 5, window size set to 5, vectors dimensions set to 100, negative sampling rate set to 10^{-3} and number of iterations set to 5). Additionally, min-count was set to 100.

Models were trained on 3 different corpora: ACL (NLP scientific articles from the ACL anthology[3]), BNC (written part of the British National Corpus[4]) and PLOS (biology scientific articles from the PLOS archive collections[5]). All corpora are the same size (about 100 million words) but they are from different types (the BNC is a generic corpus while PLOS and ACL are specialized corpora) and different domains. Corpora were lemmatized and POS-tagged using the Talismane toolkit (Urieli, 2013). Every word is associated to its POS for all subsequent experiments.

For each corpus, we trained 5 models using the exact same hyperparameters mentioned above; they only differ because of the inherent randomness of word2vec's technique. We then made 10 pairwise comparisons of models per corpus, computing the variation score for every word.

Corpus	Voc. size	Mean variation	Std. dev. (models)	Std. dev. (words)
ACL	22 292	0.16	0.04	0.08
BNC	27 434	0.17	0.04	0.08
PLOS	31 529	0.18	0.05	0.09

Table 1: Mean variation score and standard deviations for each corpus (5 models trained per corpus).

3 Models Variation

Table 1 reports the results of the comparisons. For each corpus we indicate the mean variation score, i.e. the variation averaged over all words and the

[2] https://code.google.com/archive/p/word2vec/

[3] Bird et al. (2008)

[4] http://www.natcorp.ox.ac.uk/

[5] https://www.plos.org/text-and-data-mining

10 pairwise comparisons. The variation is very similar from one corpus to the other. Standard deviation is low (average of 0.04) across the 10 pairs of models, meaning that the variation is equally distributed among the comparisons made for each corpus. The standard deviation across words is much higher (average of 0.08), which indicates that there are important differences in variation from one word to the other within the same category of models.

Variation scores for a given word can be zero (all 25 nearest neighbors are identical, although their order can vary) or as high as 0.68 (only a third of the nearest neighbors are found in both models). Based on the average variation score across the 5 models, we had a closer look at words varying the most and the least in each corpus. We identified semantic clusters that remained stable across models. E.g., in the BNC that was the case for temporal expressions (*am*, *pm*, *noon*). For all 3 corpora we identified closed classes of co-hyponyms, e.g. family members in the BNC (*wife*, *grandmother*, *sister*...), linguistic preprocessing in ACL (*parsing*, *lemmatizing*, *tokenizing*...) and antibiotics in PLOS (*puromycin*, *blasticidin*, *cefotaxime*...). For ACL and PLOS we also noticed that words belonging to the transdisciplinary scientific lexicon remained stable (conjunctive adverbs such as *nevertheless*, *moreover*, *furthermore* and scientific processes such as *hypothethize*, *reason*, *describe*). Among words displaying high variation we found a large number of tagging errors and proper nouns. We also identified some common features for other words displaying a high variation. E.g. highly polysemic words (*sign* in ACL, *make* in the BNC) and generic adjectives, i.e. adjectives than can modifiy almost any noun (*special* in ACL, *current* in PLOS and *whole* in the BNC), tend to vary more.

As there seems to be some common features of words that show a similar level of stability, we decided to try to predict the variation score.

4 Predicting the Variation

The predictive statistical models we trained are based on a set of features calculated for each word in a given distributional model. The target value is the average variation score measured across the 5 models (and 10 pairwise comparisons), so that the statistical model focuses on predicting the stability of an embedding based on a single distributional model, without having to actually train several models with the same hyperparameters. Of course, we also wanted to identify more precisely the features of stable and unstable word embeddings.

4.1 Selected Features

We measured the following features that are intrinsic to the word, corpus or model:

- *pos*: part of speech (nouns, adjectives, adverbs, verbs, proper nouns);
- *polysemy*: degree of polysemy of the word, according to an external resource;
- *frequency*: frequency of the word in the corpus;
- *entropy*: dispersion of the contexts of a word;
- *norm*: L2-norm of the vector of the word in the semantic space;
- *NN-sim*: cosine similarity of the word nearest neighbor.

POS is a straightforward feature, given by the tagger used to preprocess the corpora. As we have seen above, words in some categories such as proper nouns seemed to show higher variation than others.

To compute the degree of polysemy of a word, we used ENGLAWI, a lexical resource built from the english Wiktionary entries (Sajous and Hathout, 2015). The degree of polysemy corresponds to the number of definitions a word has in this resource. If a word does not exist in the resource, we assigned it a degree of polysemy of 1. As word embeddings aggregate all the senses of a word in a single vectors, it can be expected that polysemous words will show more variation.

Frequency of a word in a corpus is of course a very important feature when assessing embeddings (Sahlgren and Lenci, 2016). It is known that words of low or high frequencies get lower results on different tasks using embeddings.

The dispersion of the contexts of a word is measured by the normalized entropy of a word's collocates computed on a symmetrical rectangular window of 5 for open classes words only. A higher value indicate a high variability in the contexts, which should also be correlated to variation.

We chose the L2-norm of a word vector in the model as a feature since Trost and Klakow (2017) found that the L2-norm of common words do not follow the general distribution of the model.

The last feature is the cosine similarity value of the word nearest neighbor in the semantic space. It is logically expected that close neighbors of a word will have a tendency to remain stable across models.

Corpus	Mean adjusted R^2 (std. dev.)
ACL	0.39 (0.0007)
BNC	0.43 (0.0102)
PLOS	0.48 (0.0006)

Table 2: Mean adjusted R^2 score for predicting the variation of a word on ACL, BNC and PLOS.

We performed a multiple linear regression with pairwise interactions. We have 5 multilinear regression models per corpus (one per distributional model), but they all target the average variation score of a word as the predicted value. We evaluated the validity of each model using the adjusted R^2 value.

4.2 Models and Results

We can see in Table 2 that we are able to predict up to 48% of the variance, with slight differences across the three corpora. Although far from an efficient prediction, these values indicate that we nevertheless captured important features that can explain the stability of embeddings.

In order to understand the impact of the different features selected to train the regression models, we followed a feature ablation approach similar to Lapesa and Evert (2017). For each word embedding model, we trained one multilinear model using all features. We then trained 6 other models by removing one feature at a time, and computed the difference (loss) of the adjusted R^2 compared to the full 6-features model. This difference can be seen as the relative importance of the ablated feature.

Figure 1 shows the impact of each feature used for training. We can see a similar global pattern for models trained on the 3 corpora with two features displaying more importance than others. The cosine similarity of the nearest neighbor has the most important impact. As shown in Figure 1 it explains around 20% of the variance. This was expected given the way we measure variation. However, it accounts for less than half of the predictive model's power, meaning that there are other important effects involved. The POS also has a high impact on the model trained. Other features have less impact on the regression models trained. This is the case of the entropy and the polysemy for all 3 corpora. The norm and frequency have a slightly different impact depending on the corpus.

To get a better understanding of the effects of each feature on the variation of a word, we analyzed the effect of features using partial effects.

We observed similar effects of the features for all 3 corpora. As we stated before, the cosine similarity score of the nearest neighbor of a word is the most important feature when predicting its variability. We found that words having a higher nearest neighbor similarity score displayed less variation. On the contrary, when the similarity score was lower, the variation was higher. It seems logical that a very close neighbor remains stable from one model to the other. However this is not a systematic behavior. Some words having very close neighbors display a high variability.

For POS, we confirm that proper nouns have a higher variation than other categories, along with nouns on a smaller scale. No differences could be found among other categories.

The norm of the vector is negatively correlated to variation: word with vectors distant from the origin show less variation. This effect was confirmed but less clear for the ACL models. This phenomenon has to be further inquired as is the overall geometry of word embeddings vector space. E.g., Mimno and Thompson (2017) have shown that embeddings trained using *word2vec* Skip-Gram are not evenly dispersed through the semantic space.

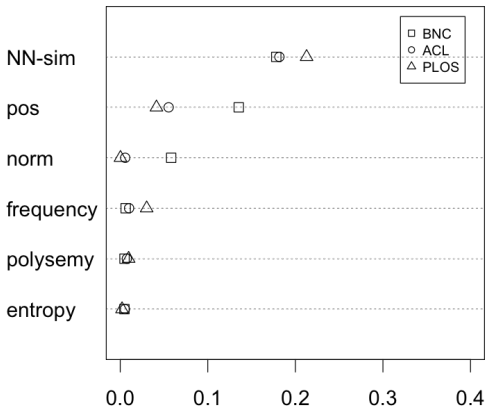

Figure 1: Feature ablation for multilinear regression models trained for ACL, BNC and PLOS.

The effect of the frequency over the predictability of the variation is not linear. Words having very low or very high frequency are more affected by variation than words in the mid-frequency range. This partly infirms the common knowledge that embeddings of more frequent words are of better quality. We actually found a number of frequent words displaying instability words in each corpus (e.g. *gene* and *protein* in PLOS, *language* in ACL and *make* in BNC etc.).

The degree of polysemy of a word also has a slight effect on the predictability of the variation of a word. The more polysemic a word is, the more likely its variation score is to be high.

As for the entropy, we observed for ACL and the BNC, that words having higher entropy with their contexts display more variation.

Concerning these two last features (polysemy and entropy) experiments confirm that distributional semantics has more difficulty in representing the meaning of words that appear in a variety of contexts.

5 Conclusion

In this paper, we wanted to get a better understanding of the intrinsic stability of neural-based word embeddings models. We agree with Antoniak and Mimno (2018) when saying that word embeddings should be used with care when used as tools for corpus linguistics, as any phenomenon observed in such models could be simply due to random.

We proposed a method that measures the variation of a word, along with a technique to predict the variation of a word by using simple features. We have seen that not all features have the same importance when predicting the variability, prominent features being the cosine similarity score of the nearest neighbor of a word and its POS. The other features we considered, while having a lesser predictive power, helped to shed some light on which areas of the lexicon are more or less affected by the variation. This means that we can hope to assess which words (in a given corpus) can be more reliably represented by embeddings, and which one should be analyzed with more caution.

Beyond the practical insterest of this prediction, this work is a step towards a better understanding of the conditions in which distributional semantics capture and represent the meaning of words. We already observed that some words or meanings are more challenging than others. In this way we assume that stability attest the quality of a semantic representation.

In this work, the embeddings models used were trained with default hyperparameters. In the future, we want to know if hyperparameters used when training word embeddings have an impact on the variation. We also want to make sure that the identified features explaining the variation will be the same when varying the hyperparameters. In the long run, this could lead to an alternative to benchmark test sets when selecting the hyperparameter values.

Acknowledgments

We are thankful to the anonymous reviewers for their suggestions and comments. Experiments presented in this paper were carried out using the OSIRIM computing platform[6] that is administered by the IRIT computer science lab and supported by the National Center for Scientific Research (CNRS), the Région Midi-Pyrénées, the French Government, and the European Regional Development Fund (ERDF).

References

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Fatemeh Torabi Asr, Jon A. Willits, and Michael N. Jones. 2016. Comparing predictive and co-occurrence based models of lexical semantics trained on child-directed speech. In *Proceedings of the 37th Meeting of the Cognitive Science Society*.

Marco Baroni, Georgiana Dinu, and German Kruszewski. 2014. 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*, pages 238–247, Baltimore, Maryland, USA.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of Language Resources and Evaluation Conference (LREC 08)*, Marrakesh, Morrocco.

[6] See http://osirim.irit.fr/site/en

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174, Berlin, Germany.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. In *ACM Transactions on Information Systems*, volume 20, page 116:131.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1489–1501, Berlin, Germany.

Johannes Hellrich and Udo Hahn. 2016. Bad company - neighborhoods in neural embedding spaces considered harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2785–2796, Osaka, Japan.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography*, 1(1):7–36.

Gabriella Lapesa and Stefan Evert. 2017. Large-scale evaluation of dependency-based DSMs: Are they worth the effort? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, Short papers, pages 394–400, Valencia, Spain.

Bofang Li, Tao Liu, Zhe Zhao, Buzhou Tang, Aleksandr Drozd, Anna Rogers, and Xiaoyong Du. 2017. Investigating different syntactic context types and context representations for learning word embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2411–2421.

Oren Melamud, David McClosky, Siddharth Patwardhan, and Mohit Bansal. 2016. The role of context types and dimensionality in learning word embeddings. In *Proceedings of NAACL-HLT 2016*, San Diego, California.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2873–2878, Copenhagen, Denmark.

Neha Nayak, Gabor Angeli, and Christopher D. Manning. 2016. Evaluating word embeddings using a representative suite of practical tasks. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP*, pages 19–23, Berlin, Germany.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531.

Kirk Roberts. 2016. Assessing the corpus size vs. similarity trade-off for word embeddings in clinical NLP. In *Proceedings of the Clinical Natural Language Processing Workshop*, pages 54–63, Osaka, Japan.

Magnus Sahlgren. 2006. *The Word-Space Model*. Ph.D. thesis, Stockholm University, Sweden.

Magnus Sahlgren and Alessandro Lenci. 2016. The effects of data size and frequency range on distributional semantic models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 975–980, Austin, Texas.

Franck Sajous and Nabil Hathout. 2015. GLAWI, a free XML-encoded machine-readable dictionary built from the French Wiktionary. In *Proceedings of the eLex 2015 conference*, pages 405–426, Herstmonceux, England.

Thomas A. Trost and Dietrich Klakow. 2017. Parameter free hierarchical graph-based clustering for analyzing continuous word embeddings. In *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing*, Vancouver, Canada.

Assaf Urieli. 2013. *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph.D. thesis, University of Toulouse, France.