Investigating the Stability of Concrete Nouns in Word Embeddings

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

Abstract

We know that word embeddings trained using neural-based methods (such as word2vec SGNS) are sensitive to stability problems and that across two models trained using the exact same set of parameters, the nearest neighbors of a word are likely to change. All words are not equally impacted by this internal instability and recent studies have investigated features influencing the stability of word embeddings. This stability can be seen as a clue for the reliability of the semantic representation of a word. In this work, we investigate the influence of the degree of concreteness of nouns on the stability of their semantic representation. We show that for English generic corpora, abstract words are more affected by stability problems than concrete words. We also found that to a certain extent, the difference between the degree of concreteness of a noun and its nearest neighbors can partly explain the stability or instability of its neighbors.

1 Introduction

Word embeddings are more and more used in corpus linguistics studies to draw conclusions on the usage of a word. Looking at a word's nearest neighbors in embeddings models is a common way to do that. Word2vec (Mikolov et al., 2013) quickly became one of the most popular tools to train word embeddings since it is easy and convenient to use and yields state of the arts results. Word2vec, like any other neural-based method, implies several random processes when preprocessing the data used for training (subsampling of the corpus) and training word embeddings (initialization of neural networks weights, dynamic window, negative sampling). As a consequence training several times using the same controllable hyperparameters (number of dimensions, window size etc.) will not yield identical models.

Although this instability is not critical when using embeddings in deep learning models for NLP applications, concerns have been raised regarding the use of word embeddings in digital humanities. Observations made by looking at nearest neighbors of a specific word might not be accurate since the nearest neighbors of a word might change from one model to the other. Recently, several studies have investigated the stability of word embeddings and the possible ways to overcome the instability triggered by word embeddings to be able to use them in digital humanities. Hellrich and Hahn (2016) studied the influence of training methods and hyperparameters on this instability. They showed that an accurate selection of the number of training epochs would help prevent the unreliability of word embeddings while preventing overfitting on the training data. Antoniak and Mimno (2018) examined the influence of the corpus used when training word embeddings and showed that embeddings are not a "a single objective view of a corpus". They also emphasized the importance of taking variability into account when observing words' nearest neighbors and showed that the size of the corpus used for training will also influence this variability, in the sense that smaller corpora trigger more variability. A good yet expensive way to overcome this variability would be to draw conclusions from several word embeddings sets trained using the same controllable parameters to confirm the observations made. Other studies investigated the influence of several features on the instability of word embeddings. Wendlandt et al.

(2018) showed that several factors contributed to the instability of word embeddings, POS being one of the most important one, and that unlike what could be expected frequency did not play a major role in the stability of word embeddings. Pierrejean and Tanguy (2018b) investigated the influence of several features that were intrinsic to a word, corpus or model on the stability of word embeddings. They proposed a technique to predict the variation of a word given simple features (POS, degree of polysemy of a word, frequency, entropy of a word with its contexts, norm of the vectors and score of the nearest neighbor of a word). They showed that the cosine similarity score of the nearest neighbor and the POS play a major role in the prediction of the variation. They also showed that words with very low or very high frequency are more affected by variation. Pierrejean and Tanguy (2018a) also identified some semantic clusters that would remain stable from one model to the other when training word embeddings using the same controllable hyperparameters. Most of these clusters seem to consist of concrete words (e.g. family members, objects and rooms of the house).

Some recent works studied the semantic representations of concrete and abstract words in count-based distributional models and showed that concrete words have concrete nearest neighbors while abstract words tend to have abstract nearest neighbors (Frassinelli et al., 2017). Naumann et al. (2018) also showed that abstract words have higher contextual variability and are thus more difficult to predict than concrete words. We wonder if those findings would also apply to models trained using neural-based methods and if different behaviors regarding variability could be observed for concrete and abstract words.

In this work we analyze the relationship between the variation of nearest neighbors of a noun and its degree of concreteness. In order to get a good understanding of this relationship, we decided to perform our analysis on 4 corpora of different sizes and types. First, we investigate the relationship existing between the variation of nearest neighbors and frequency. Then we investigate the impact of the degree of concreteness of a noun on the stability of its semantic representation. Finally, we analyze the stability of the nearest neighbors of nouns through their degree of concreteness.

2 Experiment setup

2.1 Models

We trained word embeddings using word2vec (Mikolov et al., 2013) on 4 corpora of different sizes and types. We used 2 generic corpora, the BNC made of about 100 million words[1], and UMBC, a web-based corpus made of about 3 billion words (Han et al., 2013). We also used 2 specialized corpora, ACL (NLP scientific papers from the ACL Anthology (Bird et al., 2008)) made of about 100 million words, and PLOS also consisting of about 100 million words (biology scientific papers gathered from the PLOS archive collections[2]). Corpora were lemmatized and POS-tagged using the Talismane toolkit (Urieli, 2013). For each corpus, we trained 5 models using the same following default hyperparameters: architecture Skip-Gram with negative sampling rate of 5, window size set to 5, vectors dimensions set to 100, subsampling rate set to 10^{-3} and number of iterations set to 5. We only considered words that appear more than 100 times.

2.2 Word-level variation

Computing the variation of nearest neighbors is an easy way to assess the quality of a semantic representation. Nearest neighbors that remain the same from one model to the other can be considered more reliable than neighbors that vary. To measure this we computed the degree of variation for the 25 nearest neighbors of a word between two models. The variation score corresponds to the ratio of nearest neighbours that do not appear in both models (without considering their rank). E.g., a variation score of 0.20 indicates than for the 25 nearest neighbors of a word in one model, 5 neighbors do not appear in the 25 nearest neighbors of the other model. We performed pairwise comparisons between the 5 trained models

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

[2] www.plos.org

for each corpus resulting in 10 comparisons per corpus. We computed the variation for a selected set of POS only: nouns, adjectives, verbs and adverbs. We then computed the mean variation for each word.

2.3 Concreteness

Following Naumann et al. (2018) we used the concreteness ratings presented by Brysbaert et al. (2014). Those ratings were collected using crowdsourcing. 40 000 words total were rated between 1 (abstract word) to 5 (concrete words). The instructions given to participants stated that the concreteness of a word is defined as "something you can experience through your senses". The resource contains 14 592 nouns that have an average concreteness score of 3.53 ($\pm$1.02).

Using this resource, each noun in each corpus was given a concreteness score. We excluded other POS since as it was noted by Frassinelli et al. (2017), it is easier to qualify the degree of concreteness of nouns. In the analyses performed using the degree of concreteness, we only considered words existing in the resource. This resulted in 8 796 nouns for the BNC, 5 288 nouns for PLOS, 19 720 nouns for UMBC and 3 899 nouns for ACL. We computed the average concreteness of these nouns for each corpus. This resulted in an average concreteness score of 3.38 for UMBC ($\pm$1.01), 3.48 for the BNC ($\pm$1.02), 3.28 for ACL ($\pm$1.01) and 3.47 for PLOS ($\pm$0.98).

3 Results

3.1 Intrinsic evaluation

To check the overall performance of our models, we ran an intrinsic evaluation using a standard evaluation test set, MEN (Bruni et al., 2013). MEN consists of 3000 pairs of words. Some words used in MEN pairs are not present in the different models vocabulary. Thus the evaluation was run on 1 176 pairs for ACL, 2 687 pairs for the BNC, 1 516 pairs for PLOS and 2 996 pairs for UMBC. We reported the average score (Spearman correlation) for the 5 models for each corpus in Table 1. We can see that results are different for all corpora. Generic corpora have higher scores (0.73 for the BNC and 0.70 for UMBC) compared to specialized corpora (0.51 for ACL and 0.57 for PLOS). This is not surprising because the type of evaluation test set we used is not really tailored for small specialized corpora such as ACL and PLOS.

We observed that the results were quite stable across models. As a side note, it is important to mention that most of the nouns in the MEN test set are concrete nouns with an average concreteness score of 4.6. This raises questions regarding the bias of intrinsic evaluation test sets. When evaluating distributional semantics models, what does it mean to focus mainly on evaluating the semantic representation of concrete words? If we consider word embeddings more particularly and the fact that they are prone to stability problems, how does the stability relate to the concreteness of a word? Are concrete words more stable than abstract words? We propose to investigate those effects in the following experiments.

Corpus	MEN score	Voc. size	Mean variation	Std. dev. (models)	Std. dev. (words)
ACL	0.51	22 292	0.16	0.04	0.08
BNC	0.73	27 434	0.17	0.04	0.08
PLOS	0.57	31 529	0.18	0.05	0.09
UMBC	0.70	184 396	0.22	0.05	0.10

Table 1: MEN score, vocabulary size, mean variation score and standard deviations for each corpus (5 models trained per corpus).

3.2 Global variation

To get an estimate of the proportion of instability in our models, we started by computing the variation of the 25 nearest neighbors for every word in each corpus.

Table 1 displays the vocabulary size for models trained for each corpus as well as the mean variation along with standard deviation. 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 twice as high (average of 0.09), which indicates that there are important differences in variation from one word to the other within the same category of models.

We wish to investigate the correlation between the variation score of a word and its degree of concreteness. A positive correlation would confirm that concrete words have a better semantic representation.

3.3 Frequency

Before looking at the impact of concreteness on variation, we need to understand the relationship between variation and frequency. As we can see in Table 2, variation and frequency are correlated in all our corpora. We see that less frequent words tend to vary less. This effect is clearer for specialized corpora. However we observed that the relation between variation and frequency is not linear with words in very low or high frequency range having a tendency to vary more than words in the mid-frequency range. This is partly in line with Sahlgren and Lenci (2016) who observed that it is more challenging for neural-based models to train good vectors for low-frequency words.

Corpus	Number of nouns	Nouns with concr. score	Correl. freq-var	Correl. freq-concr.	Correl. var-concr.
ACL	5 534	3 899	-0.42	-0.12	+0.10
BNC	10 266	8 796	-0.15	+0.03	-0.16
PLOS	9 751	5 288	-0.26	-0.07	+0.01 (ns)
UMBC	49 141	19 720	-0.27	-0.07	-0.16

Table 2: Spearman correlation scores between frequency and variation, frequency and degree of concreteness and variation and degree of concreteness. All correlations scores are significant at the 0.05 level except for the one where ns is indicated.

3.4 Concreteness and variation

We wanted to know if the degree of concreteness of a noun also has an impact on the variation of its nearest neighbors. We first wanted to confirm that distributionally similar words have similar concreteness scores (Frassinelli et al., 2017). To do so we selected the 1000 most concrete and 1000 most abstract nouns in the BNC. For each noun, we computed the average concreteness score of nearest neighbors that were nouns amongst its 25 nearest neighbors. We found that neighbors of concrete nouns had an average concreteness score of 4.6 and nearest neighbors of abstract nouns had an average concreteness score of 2.37 meaning that distributionally similar words do have similar concreteness scores.

We then investigated the correlation between the frequency of a noun and its degree of concreteness. As we can see in Table 2 the effect of frequency is almost null for the BNC. However we observe a weak negative correlation for ACL, PLOS and UMBC with less frequent words being more concrete.

We computed the correlation between the variation score of nouns and their degree of concreteness. We reported the results in Table 2. We observed different behaviors for the different corpora. We found that abstract words have a clear tendency to vary more in generic corpora (BNC and UMBC) with a Spearman correlation of -0.16. In the BNC, words such as *kitchen*, *wife*, *sitting-room* or *grandmother* are concrete and have a low variation score. These words correspond to the clusters identified by Pierrejean and Tanguy (2018a). On the other side of the spectrum, we found words like *legacy*, *realization*, *succession* or *coverage* that are abstract and whose neighbors vary significantly.

Things are very different for specialized corpora. While the effect is not visible in PLOS, the opposite effect is observed in ACL with a positive correlation. Concrete words such as *carrot*, *turtle*, *umbrella* or *horse* vary a lot. These words have a low frequency in the corpus (around 100 occurrences) and correspond to words that are used in examples. This also explains the higher negative correlation between

concreteness and frequency in ACL. Abstract words in ACL correspond to words that are very stable across the different models, e.g. *recall*, *precision* or *pre-processing*.

This difference in behavior observed between specialized and non-specialized corpora is not surprising since we use a resource where concreteness was defined as something you can experience through your senses. This raises questions concerning the notion of concreteness and what it means for a word to be concrete in a specialized corpus. It seems very important to consider the nature of the corpus when performing this type of experiments and to take into consideration that changing corpus equals to changing world. This question is especially crucial when working in specialized domains where the quantity of available data might be limited.

3.5 Concreteness of nearest neighbors and stability

We saw that nearest neighbors of a word tend to have a similar degree of concreteness. As we mentioned before, for a given word, amongst its nearest neighbors some will remain stable from one model to the other while others will vary. Here we propose to investigate the interaction between the degree of concreteness of the nearest neighbor of a noun and its stability. For the following experiments we chose to focus only on the BNC.

First for each noun we retrieved the union of its 25 nearest neighbors in the 5 models trained for each corpus. We only kept nearest neighbors that were nouns. For each nearest neighbor we retrieved its degree of concreteness when available in the resource we previously used. We also computed its cosine similarity with the target word in each model. We then computed the absolute difference between the degree of concreteness of the target and the degree of concreteness of the given neighbor as well as the standard deviation of the cosine scores across the 5 different models.

Then for each concrete noun (with a degree of concreteness above 4.2) we computed the Spearman correlation between the absolute difference of concreteness and the standard deviation of the cosines.

For the BNC, we found that in 65% of the cases where the correlation is significant the correlation is positive. This means that the higher the difference between the concreteness score of a target word with one of its neighbor, the more likely this neighbor is to change from one model to the other. For example, for the noun *telescope* (concr. = 5), two close neighbors like *wavelength* (concr. = 3.35) and *lens* (concr. = 4.64) have very similar average cosine scores with *telescope* (0.67 and 0.62 resp.). However the similarity score of *telescope* with the more abstract neighbor of the two (*wavelength*) displays much more variation across the 5 models (0.013 and 0.005 resp.).

This effect is less visible for more abstract target nouns. Amongst the significant correlations the positive ones are always more frequent but their proportion is lower for abstract words (down to 52% positive correlations).

4 Conclusion

We further explored the relation between concreteness and word embeddings. We already knew that concrete words have concrete nearest neighbors. We found that concrete words also present less instability problems and frequency by itself does not explain this phenomenon. Similarly, abstract words show more variation in their neighbors across distributional models. This indicates that word embeddings are more reliable for concrete words. Interestingly, evaluation test sets such as MEN consist mainly of concrete words.

Further investigations are required to fully understand the influence of the degree of concreteness of words. However, we can state that for extremely concrete words, nearest neighbors having a similar degree of concreteness with their target are the most stable.

The above results were found only in generic corpora (BNC and UMBC) and we observed the opposite effect – or no effect at all – in specialized corpora (ACL and PLOS). This is partly due to the fact that the way concreteness is defined in generic resources is not relevant for specialized corpora. Even though this work provided several elements to better understand the stability of word embeddings, we still need to investigate factors influencing the stability of word embeddings as well as their reliability.

References

Antoniak, M. and D. Mimno (2018). Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics 6*, 107–119.

Bird, S., R. Dale, B. Dorr, B. Gibson, M. Joseph, M.-Y. Kan, D. Lee, B. Powley, D. Radev, and Y. 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)*.

Bruni, E., N.-K. Tran, and M. Baroni (2013). Multimodal Distributional Semantics. *Journal of Artificial Intelligence Research 49*, 1–47.

Brysbaert, M., A. B. Warriner, and V. Kuperman (2014). Concreteness ratings for 40 thousand generally known Enlish word lemmas. *Behavior Research Methods 46*, 904–911.

Frassinelli, D., D. Naumann, J. Utt, and S. Schulte im Walde (2017). Contextual Characteristics of Concrete and Abstract Words. In *IWCS 2017 - 12th International Conference on Computational Semantics - Short papers*.

Han, L., A. L. Kashyap, T. Finin, J. Mayfield, and J. Weese (2013). UMBC EBIQUITY-CORE: Semantic Textual Similarity Systems. In *Proc. 2nd Joint Conference on Lexical and Computational Semantics, Association for Computational Linguistics*.

Hellrich, J. and U. 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*, Osaka, Japan, pp. 2785–2796.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean (2013). Distributed Representations of Words and Phrases and their Compositionality. *arXiv:1310.4546 [cs, stat]*.

Naumann, D., D. Frassinelli, and S. Schulte im Walde (2018). Quantitative Semantic Variation in the Contexts of Concrete and Abstract Words. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM)*, New Orleans, pp. 76–85.

Pierrejean, B. and L. Tanguy (2018a). Étude de la reproductibilité des word embeddings : repérage des zones stables et instables dans le lexique. In *TALN*, Rennes, France.

Pierrejean, B. and L. Tanguy (2018b). Predicting word embeddings variability. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM)*, New Orleans, pp. 154–159.

Sahlgren, M. and A. 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*, Austin, Texas, pp. 975–980.

Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Ph. D. thesis, Université Toulouse-II Le Mirail.

Wendlandt, L., J. K. Kummerfeld, and R. Mihalcea (2018). Factors Influencing the Surprising Instability of Word Embeddings. In *Proceedings of NAACL-HLT 2018*, New Orleans, pp. 2092–2102.