The Influence of Down-Sampling Strategies on SVD Word Embedding Stability

Johannes Hellrich **Bernd Kampe** **Udo Hahn**

{firstname.lastname}@uni-jena.de

Jena University Language & Information Engineering (JULIE) Lab

Friedrich-Schiller-Universität Jena, Jena, Germany

julielab.de

Abstract

The stability of word embedding algorithms, i.e., the consistency of the word representations they reveal when trained repeatedly on the same data set, has recently raised concerns. We here compare word embedding algorithms on three corpora of different sizes, and evaluate both their stability and accuracy. We find strong evidence that down-sampling strategies (used as part of their training procedures) are particularly influential for the stability of SVD_{PPMI}-type embeddings. This finding seems to explain diverging reports on their stability and lead us to a simple modification which provides superior stability as well as accuracy on par with skip-gram embeddings.

1 Introduction

Word embedding algorithms implement the latest form of distributional semantics originating from the seminal work of Harris (1954) or Rubenstein and Goodenough (1965). They generate dense vector space representations for words based on co-occurrences within a context window. They sample word-context pairs, i.e., typically two co-occurring tokens, from a corpus and use these to generate vector representations of words and their context. Changes to the algorithm's sampling mechanism can lead to new capabilities, e.g., processing dependency information instead of linear co-occurrences (Levy and Goldberg, 2014a), or increased performance, e.g., using word association values instead of raw co-occurrence counts (Bullinaria and Levy, 2007).

Word embedding algorithms commonly down-sample contexts to lessen the impact of high-frequency words (termed 'subsampling' in Levy et al. (2015)) or increase the relative importance of words closer to the center of a context window (called 'dynamic context window' in Levy et al. (2015)). The effect of using such down-sampling strategies on accuracy in word similarity and analogy tasks was explored in several papers (e.g., Levy et al. (2015)).

However, down-sampling and details of its implementation also have major effects on the stability of word embeddings (also known as 'reliability'), i.e., the degree to which models trained independently on the same data agree on the structure of the resulting embedding space. This problem has lately raised severe concerns in the word embedding community (e.g., Hellrich and Hahn (2016b); Antoniak and Mimno (2018); Wendlandt et al. (2018)) and is also of interest to the wider machine learning community due to the influence of probabilistic—and thus unstable—methods on experimental results (Reimers and Gurevych, 2017; Henderson et al., 2018), as well as replicability and reproducibility (Ivie and Thain, 2018, pp. 63:3–4).

Stability is critical for studies examining the underlying semantic space as a more advanced form of corpus linguistics, e.g., tracking lexical change (Kim et al., 2014; Kulkarni et al., 2015; Hellrich et al., 2018). Unstable word embeddings can lead to serious problems in such applications, as interpretations will depend on the luck of the draw. This might also affect high-stake fields like medical informatics where patients could be harmed as a consequence of misleading results (Coiera et al., 2018).

In the light of these concerns, we here evaluate down-sampling strategies by modifying the SVD_{PPMI} (Singular Value Decomposition of a Positive Pointwise Mutual Information matrix; Levy et al. (2015)) algorithm and comparing its results with those of two other embedding algorithms, namely, GLOVE (Pennington et al., 2014) and SGNS (Mikolov et al., 2013a,c). Our analysis is based on three corpora of different sizes and investigates effects on both accuracy and stability.

Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP, pages 18–26

The inclusion of accuracy measurements and the larger size of our training corpora exceed prior work. We show how the choice of down-sampling strategies, a seemingly minor detail, leads to major differences in the characterization of SVD_{PPMI} in recent studies (Hellrich and Hahn, 2017; Antoniak and Mimno, 2018). We also present $\text{SVD}_{\text{wPPMI}}$, a simple modification of SVD_{PPMI} that replaces probabilistic down-sampling with weighting. What, at first sight, appears to be a small change leads, nevertheless, to an unrivaled combination of stability and accuracy, making it particularly well-suited for the above-mentioned corpus linguistic applications.

2 Computational Methodology

2.1 Measuring Stability

Measuring word embedding stability can be linked to older research comparing distributional thesauri (Salton and Lesk, 1971) by the most similar words they contain for particular anchor words (Weeds et al., 2004; Padró et al., 2014). Most stability experiments focused on repeatedly training the *same* algorithm on one corpus (Hellrich and Hahn, 2016a,b, 2017; Antoniak and Mimno, 2018; Pierrejean and Tanguy, 2018; Chugh et al., 2018), whereas Wendlandt et al. (2018) quantified stability by comparing word similarity for models trained with *different* algorithms. We follow the former approach, since we deem it more relevant for ensuring that study results can be replicated or reproduced.

Stability can be quantified by calculating the overlap between sets of words considered most similar in relation to pre-selected anchor words. Reasonable metrical choices are, e.g., the Jaccard coefficient (Jaccard, 1912) between these sets (Antoniak and Mimno, 2018; Chugh et al., 2018), or a percentage based coefficient (Hellrich and Hahn, 2016a,b; Wendlandt et al., 2018; Pierrejean and Tanguy, 2018). We here use $j@n$, i.e., the Jaccard coefficient for the n most similar words. It depends on a set M of word embedding models, m, for which the n most similar words (by cosine) from a set A of anchor words, a, as provided by the 'most similar words' function $\text{msw}(a, n, m)$, are compared:

$$j@n := \frac{1}{|A|} \sum_{a \in A} \frac{|\bigcap_{m \in M} \text{msw}(a, n, m)|}{|\bigcup_{m \in M} \text{msw}(a, n, m)|} \quad (1)$$

2.2 SVD_{PPMI} Word Embeddings

The SVD_{PPMI} algorithm from Levy et al. (2015) generates word embeddings in a three-step process. First, a corpus is transformed to a word-context matrix listing co-occurrence frequencies. Next, the frequency-based word-context matrix is transformed into a word-context matrix that contains word association values. Finally, singular value decomposition (SVD; Berry (1992); Saad (2003)) is applied to the latter matrix to reduce its dimensionality and generate word embeddings.

Each token from the corpus is successively processed in the first step by recording co-occurrences with other tokens within a symmetric window of a certain size. For example, in a token sequence $\ldots, w_{i-2}, w_{i-1}, w_i, w_{i+1}, w_{i+2}, \ldots$, with w_i as the currently modeled token, a window of size 1 would be concerned with w_{i-1} and w_{i+1} only. Down-sampling as described by Levy et al. (2015) increases accuracy by ignoring certain co-occurrences while populating the word-context matrix (further details are described below). A word-context matrix is also used in GLOVE, whereas SGNS directly operates on sampled co-occurrences in a streaming manner.

Positive pointwise mutual information (PPMI) is a variant of pointwise mutual information (Fano, 1961; Church and Hanks, 1990), independently developed by Niwa and Nitta (1994) and Bullinaria and Levy (2007). PPMI measures the ratio between observed co-occurrences (normalized and treated as a joint probability) and the expected co-occurrences (based on normalized frequencies treated as individual probabilities) for two words i and j while ignoring all cases in which the observed co-occurrences are fewer than the expected ones:

$$PPMI(i,j) := \begin{cases} 0 & \text{if } \frac{P(i,j)}{P(i)P(j)} < 1 \\ log(\frac{P(i,j)}{P(i)P(j)}) & \text{otherwise} \end{cases} \quad (2)$$

Truncated SVD reduces the dimensionality of the vector space described by the PPMI word-context matrix M. SVD factorizes M in three special[1] matrices, so that $M = U\Sigma V^T$. Entries of Σ are ordered by their size, allowing to infer the relative importance of vectors in U and V. This can be used to discard all but the highest d values

[1] U and V are orthogonal matrices containing so called singular vectors. Σ is a diagonal matrix containing singular values.

and corresponding vectors during truncated SVD, so that $M_d = U_d \Sigma_d V_d^T \approx M$. Both GLOVE and SGNS start with randomly initialized vectors of the desired dimensionality d and have thus no comparable step in their processing pipeline. However, Levy and Goldberg (2014c) showed SGNS to perform as an approximation of SVD applied to a PPMI matrix.

2.3 Down-sampling

Down-sampling by some factor requires both a formal expression to define the factor, as well as a strategy to perform down-sampling according to this factor—data can either be sampled probabilistically or weighted (see below). The following set of formulae is shared by SGNS and SVD_{PPMI}, whereas GLOVE uses a distinct one.

Distance-based down-sampling depends on the distance between the currently modeled token w_i and a second token w_j in a token sequence (such as the above example). The distance d between w_i and w_j is given as:

$$d(w_i, w_j) := |j - i| \tag{3}$$

To increase the effect of the nearest—and thus assumedly most salient—tokens both SVD_{PPMI} and SGNS down-sample words based on this distance with a distance factor, df (s being the size of the window used for sampling):

$$df(w_i, w_j) := \frac{s + 1 - d(w_i, w_j)}{s} \tag{4}$$

To limit the effect of high-frequency words—likely to be function words—both algorithms also down-sample words according to a frequency factor (ff), which compares each token's relative frequency $r(w)$ with a threshold t:

$$ff(w) := \begin{cases} \sqrt{t/r(w)} & \text{if } r(w) > t \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

The frequency down-sampling factor for the co-occurrence of two tokens w_i and w_j is then given by the product of their down-sampling factors, i.e., the probabilities are treated as being independent:

$$ff(w_i, w_j) := ff(w_i) \cdot ff(w_j) \tag{6}$$

The strategy used to apply these down-sampling factors can affect accuracy and, especially, stability, as can the decision not to apply them at all. These down-sampling processes can either be probabilistic, i.e., each word-context pair is processed with a probability given by $df(w_i, w_j) \cdot ff(w_i, w_j)$, or operate by weighting, i.e., for each observed co-occurrence only a fraction of a count according to the product of df and ff is added to the word-context matrix. SGNS uses probabilistic down-sampling, GLOVE uses weighting and SVD_{PPMI} by Levy et al. (2015) allows for probabilistic down-sampling or no down-sampling at all. As SVD itself is non-probabilistic[2] (Saad, 2003, chs. 6.3 & 7.1) any instability observed for SVD_{PPMI} must be caused by its probabilistic down-sampling. We thus suggest $\text{SVD}_{\text{wPPMI}}$, i.e., SVD of a PPMI matrix with weighted entries, a simple modification which uses fractional counts according to $df(w_i, w_j) \cdot ff(w_i, w_j)$. As shown in Section 5, this modification is beneficial for both accuracy and stability.

3 Corpora

The corpora used in most stability studies are relatively small. For instance, the largest corpus in Antoniak and Mimno (2018) contains 15M tokens, whereas the corpus used by Hellrich and Hahn (2017) and the largest corpus from Wendlandt et al. (2018) each contain about 60M tokens. Pierrejean and Tanguy (2018) used three corpora of about 100M words each. Two exceptions are Hellrich and Hahn (2016a,b) using relatively large Google Books Ngram corpus subsets (Michel et al., 2011) with 135M to 4.7G n-grams, as well as Chugh et al. (2018) who investigated the influence of embedding dimensionality on stability based on three corpora with only 1.2–2.6M tokens.[3]

We used three different English corpora as training material: the 2000s decade of the Corpus of Historical American English (COHA; Davies (2012)), the English News Crawl Corpus (NEWS) collected for the 2018 WMT Shared Task[4] and a Wikipedia corpus (WIKI).[5] COHA contains 14k texts and 28M tokens, NEWS 27M texts and 550M tokens, and WIKI 4.5M texts and 1.7G tokens, respectively. COHA was selected as it is commonly used in corpus linguistic studies, whereas NEWS and WIKI serve to gauge the performance of all algorithms in general applica-

[2] Assuming that a non-stochastic SVD algorithm (Halko et al., 2011) is used, as in Levy et al. (2015).

[3] Size information from personal communication.

[4] statmt.org/wmt18/translation-task.html

[5] To ease replication, we used a pre-compiled 2014 Wikipedia corpus: linguatools.org/tools/corpora/wikipedia-monolingual-corpora/

tions. The latter two corpora are far larger than common in stability studies, making our study the largest-scale evaluation of embedding stability we are aware of.

All three corpora were tokenized, transformed to lower case and cleaned from punctuation. We used both the corpora as-is, as well as independently drawn random subsamples (see also Hellrich and Hahn (2016a); Antoniak and Mimno (2018)) to simulate the arbitrary content selection in most corpora—texts could be removed or replaced with similar ones without changing the overall nature of a corpus, e.g., Wikipedia articles are continuously edited. Subsampling allows us to quantify the effect of this arbitrariness on the stability of embeddings, i.e., how consistently word embeddings are trained on variations of a corpus. Subsampling was performed on the level of the constituent texts of each corpus, e.g., individual news articles. For a corpus with n texts we drew n samples with replacement. Texts could be drawn multiple times, but only one copy was kept, reducing corpora to $1 - 1/e \approx 2/3$ of their original size.

4 Experimental Set-up

We compared five algorithm variants: GLOVE, SGNS, SVD_{PPMI} without down-sampling, SVD_{PPMI} with probabilistic down-sampling, and $\text{SVD}_{\text{WPPMI}}$. While we could use SGNS[6] and GLOVE[7] implementations directly, we had to modify SVD_{PPMI}[8] to support the weighted sampling used in $\text{SVD}_{\text{WPPMI}}$. As proposed by Antoniak and Mimno (2018), we further modified our SVD_{PPMI} implementation to use random numbers generated with a non-fixed seed for probabilistic down-sampling. A fixed seed would benefit reliability, but also act as a bias during all analyses—seed choice has been shown to cause significant differences in experimental results (Henderson et al., 2018).

Down-sampling strategies for *df* and *ff* can be chosen independently of each other, e.g., using probabilistic down-sampling for *df* together with weighted down-sampling for *ff*. However, we decided to use the same down-sampling strategies, e.g., weighting, for both factors, taking into account computational limitations as well as results from pre-tests that revealed little benefit of mixed strategies.[9]

We trained ten models for each algorithm variant and corpus.[10] In the case of subsampling, each model was trained on one of the independently drawn samples. Stability was evaluated by selecting the 1k most frequent words in each non-bootstrap subsampled corpus as anchor words and calculating j@10 (see Equation 1).[11]

Following Hellrich and Hahn (2016a,b), we did not only investigate stability, but also the accuracy of our models to gauge potential trade-offs. We measured the Spearman rank correlation between cosine-based word similarity judgments and human ones with four psycholinguistic test sets, i.e., the two crowdsourced test sets MEN (Bruni et al., 2012) and MTurk (Radinsky et al., 2011), the especially strict SimLex-999 (Hill et al., 2014) and the widely used WordSim-353 (WS-353; Finkelstein et al. (2002)). We also measured the percentage of correctly solved analogies (using the multiplicative formula from Levy and Goldberg (2014b)) with two test sets developed at Google (Mikolov et al., 2013a) and Microsoft Research (MSR; Mikolov et al. (2013b)).

5 Experimental Results

Table 1 shows the accuracy and stability for all tested combinations of algorithm and corpus variants. Accuracy differences between test sets are in line with prior observations and general

[6] github.com/tmikolov/word2vec

[7] github.com/stanfordnlp/GloVe

[8] github.com/hellrich/hyperwords – See also further experimental code: github.com/hellrich/embedding_downsampling_comparison

[9] The strongest counterexample is a combination of probabilistic down-sampling for *df* and weighting for *ff* which lead to small, yet significant improvements in the MEN (0.703 ± 0.001) and MTurk (0.568 ± 0.015) similarity tasks (cf. Table 1). However, other accuracy tasks showed no improvements and the stability of this approach (0.475 ± 0.001) was far closer to SVD_{PPMI} with fully probabilistic down-sampling than to the perfect stability of $\text{SVD}_{\text{WPPMI}}$.

[10] Hyperparameters roughly follow Levy et al. (2015). We used symmetric 5 word context windows for all models as well as frequent word down-sampling thresholds of 100 (GLOVE) and 10^{-4} (others). Default learning rates and numbers of iterations were used for all models. Eigenvalues as well as context vectors were ignored for SVD_{PPMI} embeddings. 5 negative samples were used for SGNS. The minimum frequency threshold was 50 for COHA, 100 for NEWS and 750 for WIKI—increased thresholds were necessary due to SVD_{PPMI}'s memory consumption scaling quadratically with vocabulary size.

[11] Stability calculation was not performed directly between all 10 models, as this would result in a single value and preclude significance tests. Instead, we generated ten j@10 values by calculating the stability of all subsets formed by leaving out each model once in a jackknife procedure.

Corpus	Algorithm	Down-sampling	Word Similarity				Analogy		Stability
			MEN	MTurk	SimLex	WS-353	Google	MSR	
COHA	SVD_{PPMI}	none	0.697	**0.582**	0.318	0.591	0.248	0.226	**1.000**
		prob.	0.689	**0.571**	0.333	0.577	0.224	0.257	0.324
		weight	**0.702**	0.551	0.351	**0.594**	**0.262**	0.277	**1.000**
	SGNS	prob.	0.642	0.560	**0.394**	0.551	0.248	**0.311**	0.288
	GLOVE	weight	0.590	0.522	0.222	0.405	0.167	0.214	0.808
COHA Subs.	SVD_{PPMI}	none	0.645	**0.537**	0.267	**0.569**	0.192	0.184	0.310
		prob.	0.632	**0.519**	0.287	0.542	0.169	0.203	0.198
		weight	**0.651**	**0.534**	0.305	**0.568**	**0.206**	0.235	0.329
	SGNS	prob.	0.551	0.486	**0.363**	0.479	0.192	**0.243**	0.091
	GLOVE	weight	0.518	0.470	0.182	0.383	0.120	0.165	**0.330**
NEWS	SVD_{PPMI}	none	0.775	0.559	0.406	0.643	0.469	0.357	**1.000**
		prob.	0.784	0.561	0.431	0.666	0.492	0.445	0.654
		weight	**0.786**	0.568	**0.435**	0.667	0.502	0.444	**1.000**
	SGNS	prob.	0.739	**0.675**	0.430	**0.672**	**0.643**	**0.553**	0.652
	GLOVE	weight	0.698	0.576	0.309	0.536	0.548	0.444	0.679
NEWS Subs.	SVD_{PPMI}	none	0.771	0.558	0.401	0.623	0.445	0.335	0.584
		prob.	0.776	0.564	0.423	0.642	0.463	0.420	0.571
		weight	**0.781**	0.567	**0.430**	**0.649**	0.476	0.421	**0.635**
	SGNS	prob.	0.734	**0.673**	0.417	**0.647**	**0.601**	**0.513**	0.452
	GLOVE	weight	0.687	0.572	0.301	0.508	0.505	0.408	0.461
WIKI	SVD_{PPMI}	none	0.731	0.510	0.353	0.715	0.432	0.246	**1.000**
		prob.	**0.747**	0.571	0.392	**0.718**	0.482	0.311	0.714
		weight	0.743	0.560	**0.393**	**0.717**	0.482	0.305	**1.000**
	SGNS	prob.	0.735	**0.659**	0.372	**0.717**	**0.669**	**0.421**	0.488
	GLOVE	weight	0.744	0.651	0.354	0.667	0.653	0.397	0.666
WIKI Subs.	SVD_{PPMI}	none	0.726	0.526	0.355	0.699	0.410	0.244	0.635
		prob.	**0.742**	0.568	**0.391**	**0.706**	0.448	0.304	0.604
		weight	0.740	0.555	**0.389**	**0.704**	0.451	0.300	**0.651**
	SGNS	prob.	0.723	**0.657**	0.364	0.686	**0.629**	**0.407**	0.501
	GLOVE	weight	0.735	0.642	0.345	0.655	0.599	0.382	0.486

Table 1: Performance of different algorithms and down-sampling strategies with models trained on corpora with and without subsampling. **Bold** values are best or not significantly different by independent t-tests (with $p < 0.05$).

performance on WIKI is roughly in-line with the data reported in Levy et al. (2015).

In general, corpus size does seem to have a positive effect on accuracy. However, for MEN, MTurk and MSR the highest values are achieved with NEWS and not with WIKI. SVD_{PPMI} variants seem to be less hampered by small training corpora, matching observations by Sahlgren and Lenci (2016). Stability is clearly positively influenced by corpus size for all probabilistic algorithm variants except GLOVE, which, in contrast, benefits from small training corpora. Also, randomly subsampling corpora has a negative effect on both accuracy and stability—this can be explained by the smaller corpus size for accuracy and the differences in training material (as subsampling was performed independently for each model) for stability.

Figure 1 illustrates the stability of all tested algorithm variants. SVD_{WPPMI} and SVD_{PPMI} without down-sampling are the only systems which achieve perfect stability when trained on non-subsampled corpora. GLOVE is the third most reliable algorithm in this scenario, except for the large WIKI corpus. Corpus subsampling decreases the stability of all algorithms, with SVD_{WPPMI} still performing better than all other alternatives. The only exception is subsampled COHA where the otherwise suboptimal GLOVE narrowly (0.330 instead of 0.329; difference significant with $p < .05$ by two-sided t-test) outperforms SVD_{WPPMI} . SVD_{WPPMI} can achieve stability values on subsampled corpora that are competitive with those for SGNS and GLOVE trained on **non**-subsampled corpora. We found standard deviations for stability to be very low, the highest being 0.01 for GLOVE trained on non-subsampled WIKI, probably due to the overlap in our jackknife procedure.

Finally, we tested[12] the overall performance of each algorithm variant by first performing a Quade test (Quade, 1979) as a safeguard against type I

[12] All tests were conducted on the averaged accuracy values of the ten individual models per corpus (both sub-sampled and as-is) and algorithm variant (as listed in Table 1). Using the models directly would have been ill-advised because of their overlapping training data (see Demšar (2006, p. 15)). Analyses on individual corpora would have resulted in insufficient samples given the pre-conditions of our tests.

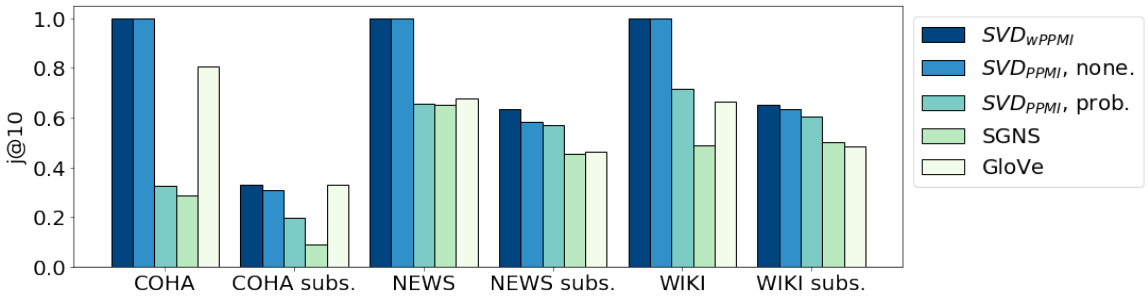

Figure 1: Stability for each combination of algorithm variant and corpus. Measured with j@10 metric (higher is better). Same data as in Table 1, standard deviations too small to display.

errors, thus confirming the existence of significant differences between algorithms ($p = 1.3 \cdot 10^{-7}$). We then used a pairwise Wilcoxon rank-sum test with Holm-Šidák correction (see Demšar (2006)) in order to compare other algorithms with $\text{SVD}_{\text{wPPMI}}$.[13] We found it to be not significantly different in accuracy from SGNS ($p = 0.101$), but significantly better than SVD_{PPMI} without down-sampling (corrected $p = 5.4 \cdot 10^{-6}$) or probabilistic down-sampling (corrected $p = 0.015$), as well as GloVe (corrected $p = 0.027$).

Our results show $\text{SVD}_{\text{wPPMI}}$ to be both highly reliable and accurate, especially on COHA, which has a size common in both stability studies and corpus linguistic applications. Diverging reports on SVD_{PPMI} stability—described as perfectly reliable in Hellrich and Hahn (2017), yet not in Antoniak and Mimno (2018)—can thus be explained by their difference in down-sampling options, i.e., no down-sampling or probabilistic down-sampling. GloVe's high stability in other studies (Antoniak and Mimno, 2018; Wendlandt et al., 2018) seems to be counterbalanced by its low accuracy and also appears to be limited to training on small corpora.

6 Discussion

We investigated the effect of down-sampling strategies on word embedding stability by comparing five algorithm variants on three corpora, two of which were larger than those typically used in stability studies. We proposed a simple modification to the down-sampling strategy used for the SVD_{PPMI} algorithm, $\text{SVD}_{\text{wPPMI}}$, which uses weighting, to achieve an otherwise unmatched combination of accuracy and stability. We also gathered evidence that GloVe lacks accuracy and is only stable when trained on small corpora.

We thus recommend using $\text{SVD}_{\text{wPPMI}}$, especially for studies targeting (qualitative) interpretations of semantic spaces (e.g., Kim et al. (2014)). Overall, SGNS provided no benefit in accuracy over $\text{SVD}_{\text{wPPMI}}$ and the latter seemed especially well-suited for small training corpora. The only downside of $\text{SVD}_{\text{wPPMI}}$ we are aware of is its relatively large memory consumption during training shared by all SVD_{PPMI} variants.

Further research could investigate the performance of $\text{SVD}_{\text{wPPMI}}$ with other sets of hyperparameters or scrutinize the effect of down-sampling strategies on the ill-understood geometry of embedding spaces (Mimno and Thompson, 2017). It would also be interesting to investigate the effect of down-sampling and stability on downstream tasks in a follow-up to Wendlandt et al. (2018).

Finally, the increasingly popular contextualized embedding algorithms, e.g., BERT (Devlin et al., 2018) or ELMo (Peters et al., 2018), are also probabilistic in nature and should thus be affected by stability problems. A direct transfer of our type specific evaluation strategy is impossible. However, an indirect one could be achieved by averaging token-specific contextualized embeddings to generate type representations.

Acknowledgments

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the *SMITH* project (grant 01ZZ1803G), Deutsche Forschungsgemeinschaft (DFG) within the *STAKI*^{2}*B*^{2} project (grant HA 2097/8-1), the SFB *AquaDiva* (CRC 1076) and the Graduate School *The Romantic Model* (GRK 2041/1).

[13] This test is a non-parametric alternative to the t-test; corrections prevent false results due to multiple comparisons.

References

Maria Antoniak and David Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–120.

M. W. Berry. 1992. Large-scale sparse singular value computations. *The International Journal of Supercomputer Applications*, 6(1):13–49.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam Khanh Tran. 2012. Distributional semantics in technicolor. In *ACL 2012 — Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers. Jeju Island, Republic of Korea, July 8–14, 2012*, pages 136–145.

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: a computational study. *Behavior Research Methods*, 39(3):510–526.

Mansi Chugh, Peter A. Whigham, and Grant Dick. 2018. Stability of word embeddings using word2vec. In *Advances in Artificial Intelligence. AI 2018 — Proceedings of the 31st Australasian Joint Conference on Artificial Intelligence. Wellington, New Zealand. December 11-14, 2018*, pages 812–818.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Enrico Coiera, Elske Ammenwerth, Andrew Georgiou, and Farah Magrabi. 2018. Does health informatics have a replication crisis? *Journal of the American Medical Informatics Association*, 25(8):963–968.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7:121–157.

Janez Demšar. 2006. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. http://arxiv.org/abs/1810.04805.

Robert M. Fano. 1961. *Transmission of Information. A Statistical Theory of Communications*. MIT Press. 3rd printing, 1966.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: the concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. 2011. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Johannes Hellrich, Sven Buechel, and Udo Hahn. 2018. JeSemE: a website for exploring diachronic changes in word meaning and emotion. In *COLING 2018 — Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. Santa Fe, NM, USA, August 20–26, 2018*, pages 10–14.

Johannes Hellrich and Udo Hahn. 2016a. An assessment of experimental protocols for tracing changes in word semantics relative to accuracy and reliability. In *LaTeCH 2016 — Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities @ ACL 2016, Berlin, Germany, August 11, 2016*, pages 111–117.

Johannes Hellrich and Udo Hahn. 2016b. Bad company: neighborhoods in neural embedding spaces considered harmful. In *COLING 2016 — Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan, December 11–16, 2016*, pages 2785–2796.

Johannes Hellrich and Udo Hahn. 2017. Don't get fooled by word embeddings: better watch their neighborhood. In *Digital Humanities 2017 — Conference Abstracts of the 2017 Conference of the Alliance of Digital Humanities Organizations (ADHO). Montréal, Quebec, Canada, August 8–11, 2017*, pages 250–252.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. Deep reinforcement learning that matters. In *AAAI-IAAI-EAAI '18 — Proceedings of the 32nd AAAI Conference on Artificial Intelligence & 30th Conference on Innovative Applications of Artificial Intelligence & 8th Symposium on Educational Advances in Artificial Intelligence. New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3207–3214.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Peter Ivie and Douglas Thain. 2018. Reproducibility in scientific computing. *ACM Computing Surveys*, 51(3):63:1–63:36.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. *New Phytologist*, XI(2):37–50. [Translation of 1901 article].

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the Workshop on Language Technologies and Computational Social Science @ ACL 2014. Baltimore, Maryland, USA, June 26, 2014*, pages 61–65.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *WWW 2015 — Proceedings of the 24th International Conference on World Wide Web: Technical Papers. Florence, Italy, May 18–22, 2015*, pages 625–635.

Omer Levy and Yoav Goldberg. 2014a. Dependency-based word embeddings. In *ACL 2014 — Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Short Papers. Baltimore, Maryland, USA, June 22–27, 2014*, pages 302–308.

Omer Levy and Yoav Goldberg. 2014b. Linguistic regularities in sparse and explicit word representations. In *CoNLL 2014 — Proceedings of the 18th Conference on Computational Natural Language Learning @ ACL 2014. Baltimore, Maryland, USA, June 26-27, 2014*, pages 171–180.

Omer Levy and Yoav Goldberg. 2014c. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27 — NIPS 2014. Proceedings of the 28th Annual Conference on Neural Information Processing Systems 2014. Montréal, Québec, Canada, December 8-13, 2014*, pages 2177–2185.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *ICLR 2013 — Workshop Proceedings of the International Conference on Learning Representations. Scottsdale, Arizona, USA, May 2–4, 2013*. https://arxiv.org/abs/1301.3781.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *NAACL-HLT 2013 — Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, GA, USA, 9–14 June 2013*, pages 746–751.

Tomáš Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013c. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 — NIPS 2013. Proceedings of the 27th Annual Conference on Neural Information Processing Systems. Lake Tahoe, Nevada, USA, December 5-10, 2013*, pages 3111–3119.

David Mimno and Laure Thompson. 2017. The strange geometry of skip-gram with negative sampling. In *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, September 7–11, 2017*, pages 2863–2868.

Yoshiki Niwa and Yoshihiko Nitta. 1994. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *COLING 1994 — Proceedings of the 15th Conference on Computational Linguistics: Volume 1. Kyoto, Japan, August 5–9, 1994*, pages 304–309.

Muntsa Padró, Marco Idiart, Aline Villavicencio, and Carlos Ramisch. 2014. Comparing similarity measures for distributional thesauri. In *LREC 2014 — Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland, May 26-31, 2014*, pages 2694–2971.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *EMNLP 2014 — Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, October 25–29, 2014*, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher T. Clark, Kenton Lee, and Luke S. Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, Louisiana, USA, June 1-6, 2018*, volume 1: Long Papers, pages 2227–2237.

Bénédicte Pierrejean and Ludovic Tanguy. 2018. Étude de la reproductibilité des word embeddings: repérage des zones stables et instables dans le lexique. In *TALN 2018 — Actes de la 25ème conférence sur le Traitement Automatique des Langues Naturelles. Rennes, France, 14-18 Mai, 2018.*, volume 1: Articles longs, articles courts de TALN, pages 33–46.

Dana Quade. 1979. Using weighted rankings in the analysis of complete blocks with additive block effects. *Journal of the American Statistical Association*, 74(367):680–683.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *WWW 2011 —*

Proceedings of the 20th International Conference on World Wide Web. Hyderabad, India, March 28 - April 1, 2011, pages 337–346.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: performance study of LSTM-networks for sequence tagging. In *EMNLP 2017 — Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, September 9-11, 2017*, pages 338–348.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Yousef Saad. 2003. *Iterative Methods for Sparse Linear Systems*, 2nd edition. Society for Industrial and Applied Mathematics, Philadelphia/PA.

Magnus Sahlgren and Alessandro Lenci. 2016. The effects of data size and frequency range on distributional semantic models. In *EMNLP 2016 — Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas, USA, November 1–5, 2016*, pages 975–980.

Gerald Salton and Michael E. Lesk. 1971. Information analysis and dictionary construction. In Gerald Salton, editor, *The* SMART *Retrieval System: Experiments in Automatic Document Processing*, chapter 6, pages 115–142. Prentice-Hall, Englewood Cliffs/NJ.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *COLING 2004 — Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland, Aug 23–27, 2004*, pages 1015–1021.

Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. In *NAACL-HLT 2018 — Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Long Papers. New Orleans, LA, USA, June 2–4, 2018*, pages 2092–2102.