Don't Get Fooled by Word Embeddings—Better Watch their Neighborhood

Johannes Hellrich
johannes.hellrich@uni-jena.de
Friedrich Schiller University Jena, Germany

Udo Hahn
udo.hahn@uni-jena.de
Friedrich Schiller University Jena, Germany

Word embeddings, such as those created by the word2vec family of algorithms (Mikolov et al., 2013), are the current state of the art for modeling lexical semantics in Computational Linguistics. They are also getting more and more popular in the Digital Humanities, especially for diachronic language research (see below). Yet the most common methods for creating word embeddings are ill-suited for deriving qualitative conclusions since they typically involve random processes that severely limit the reliability of results—repeated experiments differ in which words are deemed most similar with each other (Hellrich and Hahn, 2016a,b). We provide a short overview of different embedding methods and demonstrate how this lack of reliability might affect the outcome of experiments. We also recommend a more recent embedding method, SVD_{PPMI} (Levy et al., 2015), which seems immune to these reliability problems and, thus, much better suited (not only) for the Digital Humanities (Hamilton et al., 2016).

Word embeddings are a form of computational distributional semantics for determining a word's meaning "from the company it keeps" (Firth, 1957, p. 11), i.e., the words it co-occurs with. The word2vec algorithms have their origin in heavily trimmed artificial neural networks. Their skip-gram negative sampling (SGNS) variant is widely used because of its high performance and robustness (Mikolov et al., 2013; Levy et al., 2015). Two other word embedding methods were inspired by word2vec: GloVe (Pennington et al., 2014) tries to avoid the opaqueness stemming from word2vec's neural network heritage through an explicit word co-occurrence table, while the more recent SVD_{PPMI} (Levy et al., 2015) is built upon the classical pointwise mutual information co-occurrence metric (Church and Hanks, 1990) enhanced with pre-processing steps and hyper-parameters from the two aforementioned algorithms.

There are two sources of randomness affecting the training of SGNS and GloVe embeddings: First, the random initialization of all word embedding vectors before any examples are processed. Second, the order in which these examples are processed. Both can be replaced by deterministic alternatives, yet this would simply replace a random distortion with a fixed one, thus providing faux reliability only useful for testing purposes. In contrast, SVD_{PPMI} is conceptually not affected by such reliability problems, as neither random initialization takes place nor is a relevant processing order established.

Word embeddings can be compared with each other to measure the similarity of words (typically by cosine)—an ability by which they are often assessed (see e.g., Baroni et al. (2014) for more details on their evaluation). In the Digital Humanities, they have already been used to directly track diachronic changes in word meaning by comparing representations of the same word at different points in time (Kim et al., 2014; Kulkarni et al., 2015; Hellrich and Hahn, 2016c; Hamilton et al., 2016). They can also be used to track clusters of similar words over time and, thus, model the evolution of topics (Kenter et al., 2015) or compare neighborhoods in embedding spaces for preselected words (Jo, 2016). Besides temporal variations, word embeddings are also suited for analyzing geographic ones, e.g., the distinction between US American and British English variants (Kulkarni et al., 2016). In most of these approaches, the local neighborhood of selected words in the resulting embedding spaces, i.e., words deemed to be most similar with a word in question, are used to approximate their meaning at a given point in time or in a specific domain. Yet the aforementioned randomness leads to a lack of replicability, since repeated experiments using the same data set and algorithms result in different neighborhoods and might thus mislead researchers.

To investigate this problem, we trained three models each with three embedding methods, i.e., GloVe and SVD_{PPMI}, on the same data set and measured how they differ in their outcomes on word neighborhoods. Our data set consists of 645 German texts from the 19th century that are part of the *Deutsches Textarchiv Kernkorpus* (DTA) [German text archive core corpus] (Geyken, 2013; Jurish, 2013). The DTA contains manually transcribed texts selected for their representativeness

and cultural importance; we use the orthographically normalized and lemmatized version, with casefolding. We evaluate the word embedding methods by calculating the percentage of neighbors for the most frequent nouns in the DTA on which all three models of each method agree. Overall, SVD_{PPMI} provides perfect reliability, while the other two embedding methods lack reliability, SGNS dramatically so, which is consistent with our prior studies on word2vec (Hellrich and Hahn, 2016a,b).

Figure 1 shows the reliability for each model evaluated against the 1000 most frequent nouns in the DTA when their first ten closest neighbors (from one up to ten) are compared. Larger neighborhood size had a small positive effect on the reliability of SGNS and GloVe, yet is clearly unable to mitigate the inherent unreliability of these methods. A small inverse effect can be observed when the number of the most frequent nouns is modified while keeping a constant neighborhood size of five, as displayed in Figure 2. Finally, Table 1 provides differing neighborhoods for *Herz* [heart] as a qualitative example. In this case, though not necessarily in general, SGNS models featured a more anatomical view (e.g., *bluten* [to bleed]), whereas GloVe models uncovered metaphorical meaning (e.g., *gemüt* [mind]) and SVD_{PPMI} came out with a mix thereof. Using SGNS or GloVe models to assess a word's meaning can be strongly misleading, as evidenced by e.g., three SGNS models representing three different runs under the same experimental set-up. They lead to completely different semantic characterizations of *Herz* [heart], since two provide negatively connotated words (e.g., *schmerzen* [pain]) as closest neighbors, whereas the third provides a more positive impression (e.g., *herzen* [to caress]).

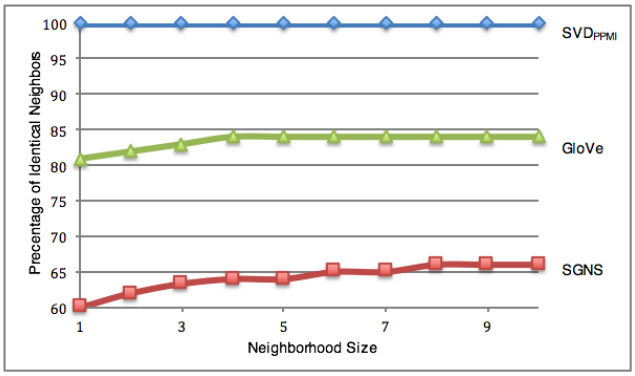

Figure 1: Reliability of different word embeddings as percentage of identical neighbors among the one to ten closest neighbor(s) to the 1000 most frequent nouns.

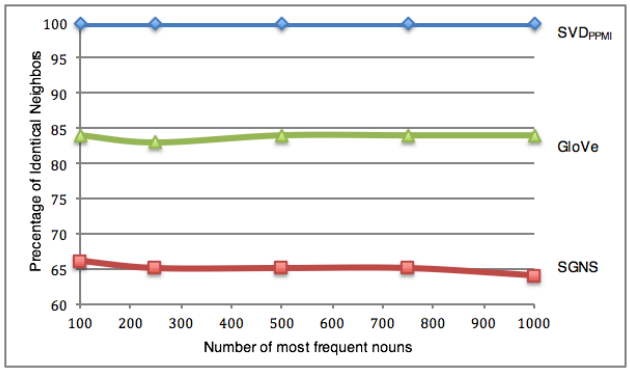

Figure 2: Reliability of different word embeddings as percentage of identical neighbors among the five closest ones for the 100 to 1000 most frequent nouns.

Embedding Model	First Neighbor	Second Neighbor	Third Neighbor	Fourth Neighbor	Fifth Neighbor
SGNS 1	~~schmerzen~~ [pain]	~~beklommen~~ [anxious]	~~busen~~ [bosom]	~~bluten~~ [to bleed]	~~herzen~~ [to caress]
SGNS 2	~~bluten~~ [to bleed]	~~klopfend~~ [beating]	~~busen~~ [bosom]	~~beklommen~~ [anxious]	~~herzen~~ [to caress]
SGNS 3	~~herzen~~ [to caress]	~~busen~~ [bosom]	~~klopfend~~ [beating]	~~beklommen~~ [anxious]	~~bluten~~ [to bleed]
GloVe 1	~~gemüt~~ [mind]	~~mein~~ [my]	~~seele~~ [soul]	~~liebe~~ [love]	~~brust~~ [chest]
GloVe 2	~~gemüt~~ [mind]	~~mein~~ [my]	~~seele~~ [soul]	~~brust~~ [chest]	~~liebe~~ [love]
GloVe 3	~~gemüt~~ [mind]	~~mein~~ [my]	~~seele~~ [soul]	~~brust~~ [chest]	~~liebe~~ [love]
SVD_{PPMI}, all	~~busen~~ [bosom]	~~fühlen~~ [to feel]	~~liebe~~ [love]	~~schmerzen~~ [pain]	~~menschenherz~~ [human heart]

Table 1: Neighborhoods for Herz [heart] as provided by different word embedding models.

The lack of reliability we observed is definitely problematic, as often, especially for illustrations, rather small neighborhoods are used to gauge a word's meaning. Our experimental data lead us to caution when SGNS or GloVe word neighborhoods are used for uncovering lexical semantics. We recommend SVD_{PPMI} instead, as its results are of similar quality yet guaranteed to be reliable (Levy et al., 2015; Hamilton et al., 2016). Consequently, we adapted our ongoing research activities on tracking language change to these insights and replaced the results of earlier work with SGNS (Hellrich and Hahn, 2016c) by data based on SVD_{PPMI} (Hellrich and Hahn, 2017).

Acknowledgements

This research was conducted within the Graduate School "The Romantic Model" supported by grant GRK 2041/1 from the Deutsche Forschungsgemeinschaft (DFG).

Bibliography

Baroni, M., Dinu, G. and Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pp. 238–47.

Church, K.W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1): 22–29.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*, pp. 1–32.

Geyken, A. (2013). Wege zu einem historischen Referenzkorpus des Deutschen: das Pro- jekt Deutsches Textarchiv. *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*, pp. 221– 34.

Hamilton, W.L., Leskovec, J. and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long Papers, pp. 1489–501.

Hellrich, J. and Hahn, U. (2016a). An assessment of experimental protocols for tracing changes in word semantics relative to accuracy and reliability. *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities @ ACL 2016,* pp. 111–7.

Hellrich, J. and Hahn, U. (2016b). Bad company—Neighborhoods in neural embedding spaces considered harmful. *Proceedings of the 26th International Conference on Computational Linguistics,* pp. 2785–96.

Hellrich, J. and Hahn, U. (2016c). Measuring the dynamics of lexico-semantic change since the German Romantic period. *Digital Humanities 2016*, pp. 545–7.

Hellrich, J. and Hahn, U. (2017). Exploring Diachronic Lexical Semantics with JeSemE. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Jo, E.S. (2016). Diplomatic history by data. Understanding Cold War foreign policy ideology using networks and NLP. *Digital Humanities 2016*, pp. 582–5.

Jurish, B. (2013). Canonicalizing the Deutsches Textarchiv. In Hafemann, I. (ed.), *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*. pp. 235–44.

Kenter, T., Wevers, M., Huijnen, P. and de Rijke, M. (2015). Ad hoc monitoring of vocabulary shifts over time. *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pp. 1191–200.

Kim, Y., Chiu, Y., Hanaki, K., Hegde, D. and Petrov, S. (2014). Temporal analysis of language through neural language models. *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pp. 61–5.

Kulkarni, V., Al-Rfou, R., Perozzi, B. and Skiena, S. (2015). Statistically significant detection of linguistic change. *Proceedings of the 24th International Conference on World Wide Web*. pp. 625–35.

Kulkarni, V., Perozzi, B. and Skiena, S. (2016). Freshman or fresher? Quantifying the geographic variation of language in online social media. *Proceedings of the 10th International AAAI Conference on Web and Social Media*, pp. 615–8.

Levy, O., Goldberg, Y. and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3: 211–25.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems 26*, pp. 3111–9.

Pennington, J., Socher, R. and Manning, C.D. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–43.