Times Are Changing: Investigating the Pace of Language Change in Diachronic Word Embeddings

Stephanie Brandl, David Lassner
Machine Learning Group, TU Berlin, Berlin, Germany
{stephanie.brandl, lassner}@tu-berlin.de

Abstract

We propose Word Embedding Networks (WEN), a novel method that is able to learn word embeddings of individual data slices while simultaneously aligning and ordering them without feeding temporal information a priori to the model. This gives us the opportunity to analyse the dynamics in word embeddings on a large scale in a purely data-driven manner. In experiments on two different newspaper corpora, the New York Times (English) and Die Zeit (German), we were able to show that time actually determines the dynamics of semantic change. However, we find that the evolution does not happen uniformly, but instead we discover times of faster and times of slower change.

1 Introduction

Vectorial representation of natural language, known as word embeddings, have been widely used in e.g. text classification (Joulin et al., 2016) and machine translation (Mikolov et al., 2013).

As in Kim et al. (2014); Kulkarni et al. (2015); Hamilton et al. (2016) and Szymanski (2017) aligned sets of embeddings have also been used to detect changes in vectorial representations of words over time. In the past, those changes have mostly been studied at the word-level.

We propose a novel method to investigate the pace of language change based on the entire embedding matrix. Previous approaches have not been able to carry out this type of analysis, as they have taken the continuous change of language for granted and investigated those dynamics in a supervised manner.

Therefore, we present Word Embedding Networks (WEN), a method that has no knowledge about the chronological order of the slices, so we can investigate semantic changes on the whole vocabulary purely data-driven and unsupervised. Pairwise relations between embeddings and the embeddings themselves are learned simultaneously without feeding the temporal information a priori into the algorithm. In that, it is substantially more flexible than those methods mentioned above. This means that dynamics between any slicing of a text corpus can be learned (especially those where there is no order known) and the result not only contains embeddings for each slice, but an order of slices that corresponds to the dynamics of word meanings.

This method also overcomes the need of a two-step solution for aligned temporal embeddings, as has also been done by Yao et al. (2018) - the two-step solution has, according to Yao et al. (2018), its weaknesses especially in the case of non-uniformly distributed amounts of data across the slices. Closer proximities between embeddings denote time intervals of slower semantic changes, embeddings are farther apart when times are changing faster.

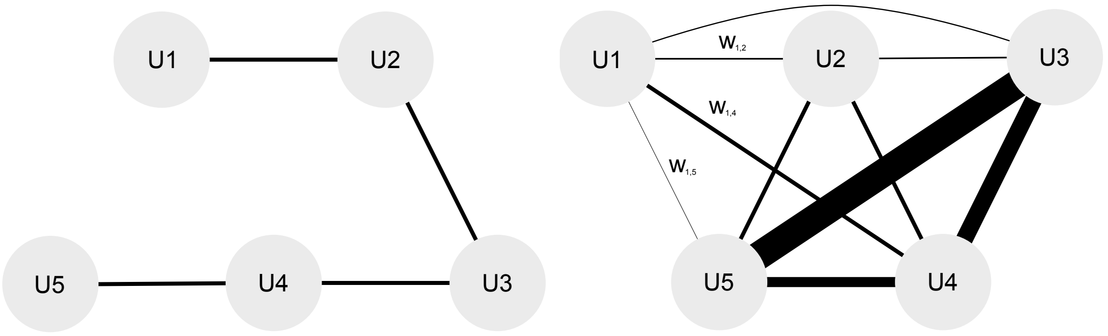
2 Related Work

Rudolph and Blei (2018) analyse dynamical changes in word embeddings based on exponential family embeddings, a probabilistic framework that generalizes the concept of word embeddings to other types data (Rudolph et al., 2016). They focus on word-level changes within and between text corpora spanning from the 19th century until today.

The authors of Yao et al. (2018) proposed a new method to learn individual word embeddings for each year of the New York Times data set (1990-2016) while simultaneously aligning the embeddings to the same vector space. Their neighborhood constraint

$$\frac{\tau}{2}\left(\|U_{t-1}-U_t\|_F^2+\|U_t-U_{t+1}\|_F^2\right)$$

Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change, pages 146–150
Florence, Italy, August 2, 2019. ©2019 Association for Computational Linguistics



(a) Dynamic Word Embeddings from (Yao et al., 2018) has a predefined ordering of embeddings $U_\cdot$.

(b) WEN learns embeddings $U_\cdot$ and $\omega_{\cdot,\cdot}$ in turn. Thicker edges denote stronger relation between embeddings.

Figure 1: Comparison of Dynamic Word Embeddings (Yao et al., 2018) and WEN which can be seen as a generalization of the former.

encourages alignment of the word embeddings. The parameter τ controls the dynamic, thus how much neighboring word embeddings are allowed to differ ($\tau = 0$: no alignment and $\tau \rightarrow \infty$: static embeddings).

3 Method

To identify the pace of change, we introduce a new method named *Word Embedding Networks* (WEN). WEN learns embeddings for e.g. different time slices while simultaneously aligning and ordering them. WEN starts with assuming an equal distance between all embeddings and then, over time, shapes the relations by moving certain embeddings closer and others farther apart. In Fig. 1 we illustrate an exemplary trained word embedding network and compare it to Dynamic Word Embeddings from Yao et al. (2018).

In order to train the weights of the graph, we include an additional weighting term $\omega_{t,t'}$ into the model and optimize over

$$\begin{aligned}\min_{U_t} F_t = \min_{U_t} \frac{1}{2} \left\| Y_t - U_t U_t^\top \right\|_F^2 \quad (1)\\ + \frac{\lambda}{2} \left\| U_t \right\|_F^2 \quad (2)\\ + \frac{\tau}{2} \sum_{t' \neq\, t}^{N} \omega_{t,t'} \left(\left\| U_t - U_{t'} \right\|_F^2 \right). \quad (3)\end{aligned}$$

Here, $U_t \in \mathbb{R}^{V \times D}$ contains the D-dimensional word embeddings in a vocabulary of size V at time point t and $Y_t \in \mathbb{R}^{V \times V}$ represents the PPMI matrix (Yao et al., 2018).

While Term 1 is responsible for training the word embeddings with respect to Y, Term 2 enforces sparse vectorial representations.

By updating $\omega_{t,t'}$ with respect to the distances between word embeddings of different slices it is meant to strengthen connections of word embeddings that lie closer together in the corresponding vector space.

To update $\omega_{t,t'}$ we first introduce a symmetric normalization function

$$\text{norm\_sym}(x_{ij}) = \frac{x_{ij}}{\left(\sum_k x_{ik} + \sum_j x_{kj} \right)}$$

where $x_{ij} \in \mathbb{R}$.

The weighting term $\omega_{t,t'}$ is then updated accordingly:

$$\begin{aligned} d_{t,t'} &= \text{norm\_sym} \left(\frac{1}{\left\| U_t - U_{t'} \right\|_F^2} \right) \\ \omega_{t,t'}^{\text{new}} &= \text{norm\_sym} \left(\omega_{t,t'} + d_{t,t'} \right). \quad (4)\end{aligned}$$

We optimize U with gradient descent. We therefore use Adam (Kingma and Ba, 2014) with default values for β and a customized learning rate (see Sec. 4.2).

We did not tune for efficiency and stopped the optimization after 1500 rounds where one round is finished when embeddings of all time slices have been updated once in a random order.

We initialize $\forall t, t'$ $\omega_{t,t'} = 1$ and update every 100 rounds according to Eq. 4.

We have implemented this in PyTorch.

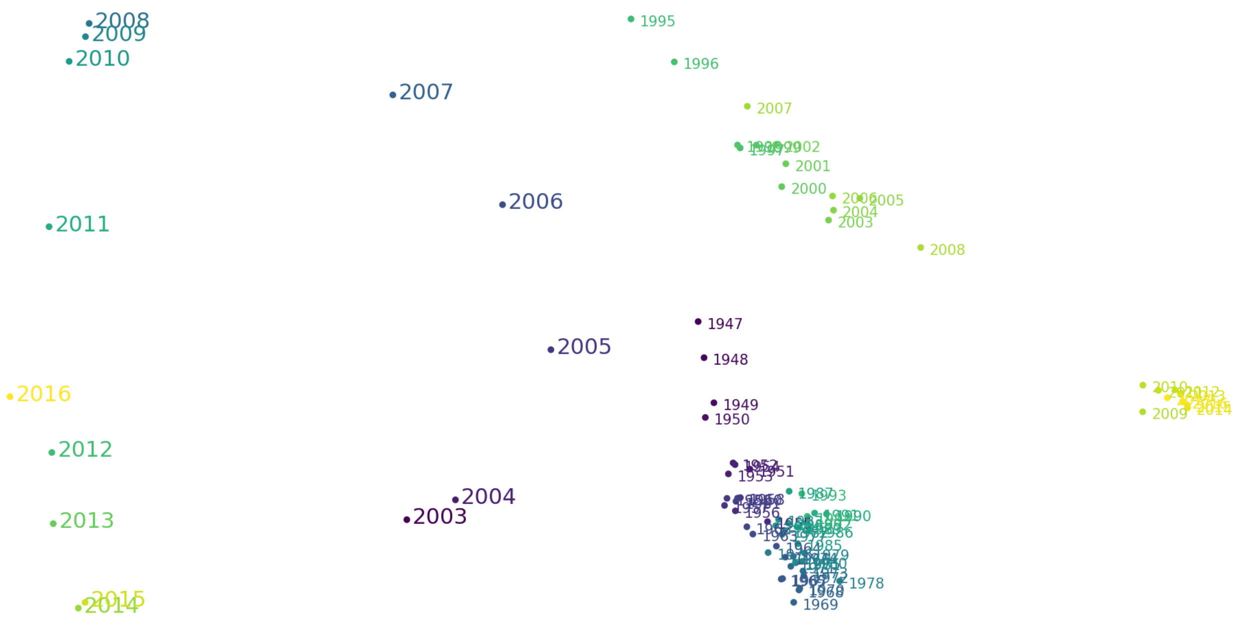
4 Training

4.1 Data Sets

New York Times 1990-2016:

The New York Times data set[1] (NYT) contains

[1] https://sites.google.com/site/zijunyaorutgers/



(a) 2-dimensional embedding of the New York Times w matrix. Slices are sorted nicely in a circular structure with only few excepions.

(b) 2-dimensional embedding of the Die Zeit w matrix. The embeddings-embedding still resembles the chronology but there is also a secondary structure of three clusters.

Figure 2: Laplacian eigenmaps of the w matrix to visualize the relationship between embeddings (pace of change). Points are colored with ground truth information.

headlines and lead texts of news articles published online and offline between 1990 and 2016 with a total of 99.872 documents.

Die Zeit 1947-2017:
Die Zeit is a German national weekly newspaper that started publishing in 1946. We obtained titles, teaser titles and teaser texts of 508.698 news articles from the Die Zeit developer API[2] that have been published online and offline between 1947 and 2017.

4.2 Parameters

We perform a grid search to select optimal parameters on the first half (1990-2002) of NYT which results in the same parameter combination as reported by (Yao et al. (2018), $\lambda = 10$, $\tau = 50$, embedding dimension= 32). We start with a learning rate of $\eta = 10^{-3}$, reducing it after 500 rounds to $\eta_{500} = 5 \cdot 10^{-4}$ and after 1000 rounds to $\eta_{1000} = 10^{-4}$.

4.3 Preprocessing

We lemmatize the data with spacy [3].
We only consider the 20.000 most frequent (lemmatized) words of the entire data set that are also under the 20.000 most frequent words in at least 3 yearly slices. This way, we filter out "trend" words that only are of significance within a very short time period. The 100 most frequent words are filtered out as stop words.

4.4 Experiments

New York Times 2003-2016:
We apply WEN with the parameters from Section 4.2 to the second part of NYT (2003-2016) and train embeddings for yearly slices (14 in total). In most of the cases, namely 85%, WEN aligns embeddings closest to each other that in fact also correspond to their chronological neighbor. We define this as the (neighborhood) accuracy of 85%.

Die Zeit 1947-2017:
We further apply WEN on the Die Zeit data set (1947-2017) to evaluate the model on a non-English text corpora which has not been involved in parameter search. We train and sort the word embeddings on the entire data set (71 yearly slices). After 1500 rounds, we achieve an accuracy of 67%.

5 Results

The high neighborhood accuracies, indeed indicate a temporal dynamic in both data sets.
To visualize the network weights as a map, w was used as an affinity matrix to generate Lapla-

[2] http://developer.zeit.de
[3] https://spacy.io

cian eigenmaps (Belkin and Niyogi, 2002) of the neighborhoods between yearly slices. In Fig. 2 maps of NYT (second half) and Die Zeit are shown side-by-side.

NYT shows a very nice circular structure of neighborhoods but also how some years lie closer together than others. For instance the years 2008-2010 can be found very close together whereas 2011 having larger gaps to its two closest neighbors 2010 and 2012. By identifying words of largest change within 2011, we found mostly personal names and companies in the high ranks whose media coverage changed during that time. Also, we detected larger gaps in actual neighboring years when there are shifts in how sections are distributed in the data set. As *Sports* remains the section with most articles throughout the entire data set, there are more documents in *Opinion* after 2003 and only half the documents in *New York and Region* after 2005.

Regarding the Die Zeit map, we observe three distinct clusters, one before 1995, one after 2008. The gaps clearly correspond to changes in either publication strategy (starting 2009 with emphasizing the online publication) and archival data storage (there are close to no teaser texts available for 1995). Both events led to a sudden change of the amounts of data available.

6 Conclusion

Word Embedding Networks (WEN) learns word embeddings of individual data slices while simultaneously aligning and ordering them in an unsupervised manner.

After being trained on news articles from the New York Times (1990-2002), the model could successfully be applied to news articles from the same corpus (2003-2016) and to data containing German newspaper articles from 1947-2017 (die Zeit). Results on both data sets show a clear temporal dynamic as 85% and 67% respectively of the closest time slices correspond to the neighboring years. Time can thus be identified as the dominant component that is governing change in word meaning in both data sets.

However, it could be shown for both data sets that change is not introduced at a constant pace hence there are times of slower and times of faster change. We found that distributional changes within the data set can have a huge influence on the perceived pace of semantic change. Therefore we argue for caution when applying models that assume continuous change, especially concerning the NYT data set, with its widespread use.

For further research, we would like to expand the experiments to corpora where the underlying slices are not ordered. For example given a corpus of works grouped by authors, we could train the model to find proximities between authors based on the similarity of meaning of the words they use.

Acknowledgments

This work was supported by the Federal Ministry of Education and Research (BMBF) for the MALT III project (01IS17058). We thank M. Alber for valuable comments on the manuscript.

References

Mikhail Belkin and Partha Niyogi. 2002. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems*, pages 585–591.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. *arXiv preprint arXiv:1405.3515*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635. International World Wide Web Conferences Steering Committee.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting Similarities among Languages for Machine Translation. *CoRR*, abs/1309.4168.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1003–1011. International World Wide Web Conferences Steering Committee.

Maja Rudolph, Francisco Ruiz, Stephan Mandt, and David Blei. 2016. Exponential family embeddings. In *Advances in Neural Information Processing Systems*, pages 478–486.

Terrence Szymanski. 2017. Temporal Word Analogies: Identifying Lexical Replacement with Diachronic Word Embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 448–453.

Zijun Yao, Yifan Sun, Weicong Ding, Nikhil Rao, and Hui Xiong. 2018. Dynamic Word Embeddings for Evolving Semantic Discovery. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 673–681. ACM.