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

Stephanie Brandl, David Lassner

Machine Learning Group, TU Berlin, Berlin, Germany

eMail: stephanie.brandl@tu-berlin.de

Introduction

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.

Related Work

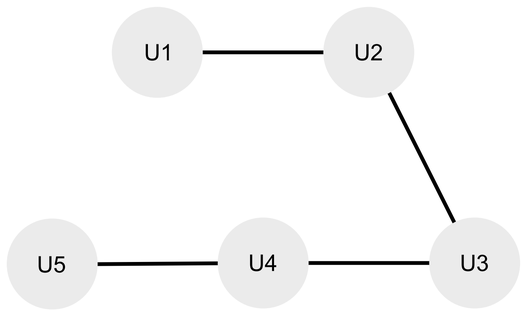
The authors of [1] analyse dynamical changes in word embeddings based on exponential family embeddings, a probabilistic framework that generalizes the concept of word embeddings to other types of data [2]. They focus on word-level changes within and between text corpora spanning from the 19th century until today.

In [3] a new method has been proposed 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

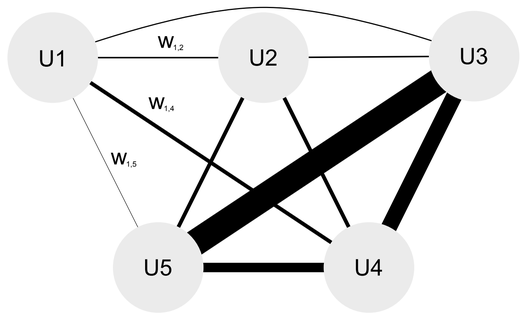
$$\frac{\tau}{2}\left(\|U_{t-1} - U_t\|_F^2 + \|U_t - U_{t+1}\|_F^2\right)$$

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 $\tau \to \infty$: static embeddings).

Method



(a) Dynamic Word Embeddings from [3] has a predefined ordering of embeddings $U_.$.



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

Figure 1: Comparison of Dynamic Word Embeddings [3] and WEN which can be seen as a generalization of the former.

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

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

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 [3].

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

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

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

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

Experiments

Data Sets

New York Times 1990-2016:

- headlines and lead texts
- 99.872 documents
- 1990-2002 for parameter selection
- 2003-2016 for experiments

Die Zeit 1947-2017

- titles, teaser titles and teaser texts
- 508.698 documents
- parameters from NYT
- 1947-2017 for experiments

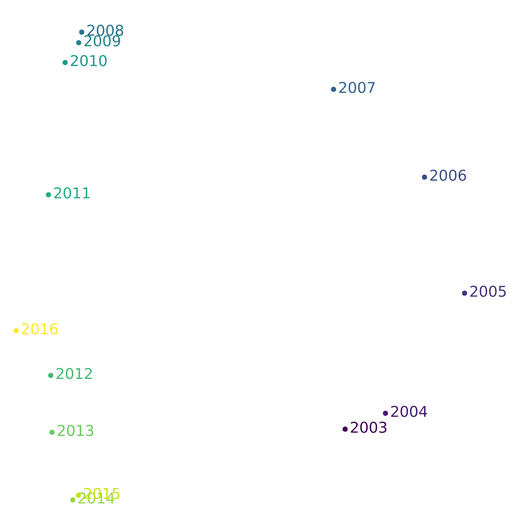
Experiments cont'd

Parameters

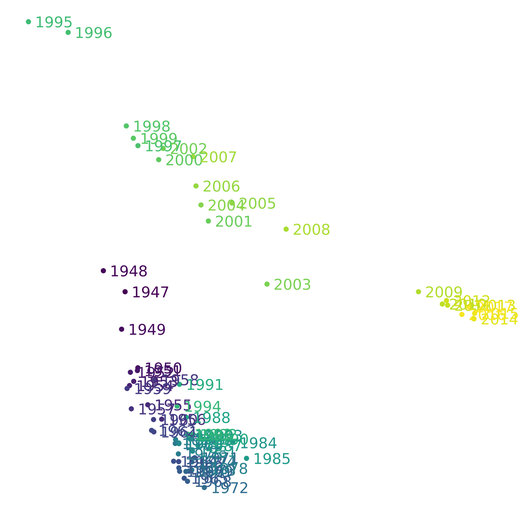
grid search on first half of NYT

- $\lambda = 10, \tau = 50, D = 32$
- $\eta = 10^{-3}, \eta_{500} = 5 \cdot 10^{-4}, \eta_{1000} = 10^{-4}$

Results



(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 embedding still resembles the chronology but there is also a secondary structure of 3 clusters.

Figure 2: Laplacian eigenmaps [4] of the affinity matrix of w, the neighborhood between yearly slices

New York Times

- larger gaps around 2011
- words of largest change within 2011 mainly companies or personal names
- larger gaps in actual neighboring years when there are shifts in how sections are distributed in the data set, see Figure 3
- **neighborhood accuracy of 85%**

Die Zeit

- three distinct clusters: $1947 - 1995, 1995 - 2008, 2008 - 2017$
- changes in publication strategy (emphasis on online publication after 2008)
- changes in archival data storage (no teaser texts for 1995)
- **neighborhood accuracy of 67%**

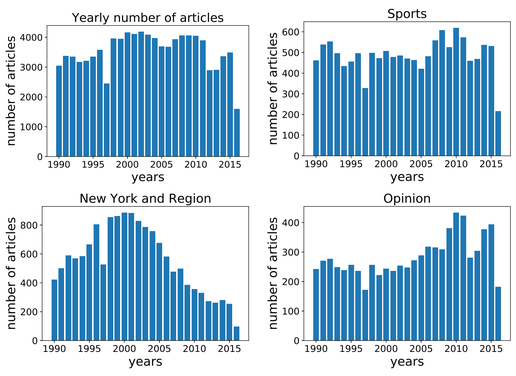


Figure 3: Statistics on the New York Times corpora

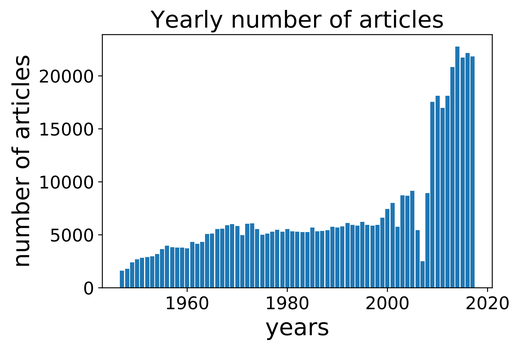


Figure 4: Statistics on the die Zeit publishing activity.

Conclusion

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

The model could successfully be applied to news articles from the New York Times (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% 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. We found that distributional changes within the data set can have a huge influence on the perceived pace of semantic change.

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.

References

[1] Rudolph, M., and Blei, D. (2018). Dynamic embeddings for language evolution. In Proceedings of the 2018 World Wide Web Conference (pp. 1003-1011). International World Wide Web Conferences Steering Committee.

[2] Rudolph, M., Ruiz, F., Mandt, S., and Blei, D. (2016). Exponential family embeddings. In Advances in Neural Information Processing Systems (pp. 478-486).

[3] Yao, Z., Sun, Y., Ding, W., Rao, N., and Xiong, H. (2018). Dynamic word embeddings for evolving semantic discovery. Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (pp. 673-681). ACM.

[4] Belkin, M. and Niyogi, P. (2002). Laplacian eigenmaps and spectral techniques for embedding and clustering. In Advances in neural information processing systems (pp. 585-591).