# NLP Project Description

Master IASD - Paris Dauphine - PSL University

Guillaume Bressan, João Paulo Casagrande Bertoldo, Oskar Rynkiewicz

13 February 2019

## 1 Overview

As part of the course *Natural Language Processing* of the IASD Master at Paris Dauphine PSL, we present a description of our project will be based on the paper *Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change* [5]. The authors use word embeddings to analyse historical semantic changes in large corpora in different languages. The original project's page with links to all its resources can be found *here*. Our project will be in *this repository*.

By training models with sub-corpora of different time periods, the authors were able to use similarity metrics in the embeddings to quantify the rate of meaning changes in the vocalary over time. They trained three different types of model: PPMI (Positive Point-wise Mutual Information), SVD (Singular Value Decomposition), and SGNS (or word2vec) [7]. Then, the embeddings' axes were aligned using orthogonal Procrustes analysis and tested on several benchmark tasks, resulting in the choice of the SGNS embedding. Finally, they fitted a linear model to quantify the rate of meaning change as a function of word frequency and polysemy.

## 2 Data

The original paper used 6 corpora from Google N-Grams [6] and COHA [2] in 4 languages (English, French, German, and Chinese). For this project, only English will be considered. Of the three datasets in English analysed in the paper, we will consider the COHA corpus.

This dataset has been made to be genre-balanced and representative of American English. Besides, it had expressive results in the paper. The authors of the reference paper worked both on a version of COHA with lemmatization and another without. For the sake of practicality, we will only use the lemmatized corpus because it is smaller.

The authors of original project made available their pre-trained embeddings, as well as historical word frequency, and other metrics (i.g. a polysemy score) used in the paper. We will try to use their pre-computed values to extent possible both to avoid long code executions and to be able to compare our the results. A detailed data description can be found *here*.

## 3 The task

This project will use the original paper's procedure as the starting point and then modify some aspects, that will be specified in the next section. Here we briefly explain what the initial task consisted of.

### Obtaining word embeddings [1]

Based on 4 benchmark tests, the authors of [5] concluded that SGNS (a.k.a. 'word2vec') is the better than PPMI and SVD for the purpose of detecting word semantic diachronic (historical) changes. Therefore, we only consider SGNS models in our project.

Each word $w_i$ is represented by a low-dimensional vector (its embedding) $\mathbf{w}_i \in \mathbb{R}^d$ and a context vector $\mathbf{c}_i \in \mathbb{R}^d$. These vectors are trained to approximate

$$\hat{p}\left(w_j|w_i\right) \propto \exp\left(\mathbf{w}_i \cdot \mathbf{c}_j\right) \tag{1}$$

[1] This corresponds to the section 2 in the paper.

where $\hat{p}\left(w_j|w_i\right)$ is the empirical probability of seeing $w_j$ in a fixed-length window of text centered on $w_i$. The corpus is splitted in chunks of 10 years and a different model is trained for each split.

As a result, we can extract, for each word and each decade $t$, an embedding $\mathbf{w}_i^{(t)}$. However, to be able to compare two embeddings $\mathbf{w}_i^{(t)}$ and $\mathbf{w}_i^{(t+1)}$ of the same word, the vector spaces need to have their axes aligned, which is done using orthogonal Procrustes analysis.

Let $\mathbf{W}^{(t)} \in \mathbb{R}^{d\times|\mathcal{V}|}$[2] be the matrix of embeddings learned at the decade $t$. The two embeddings are aligned by optimizing:

$$\mathbf{R}^{(t)} = \underset{\mathbf{Q}^\top\mathbf{Q}=\mathbf{I}}{\operatorname{argmin}} \left\| \mathbf{Q}\mathbf{W}^{(t)} - \mathbf{W}^{(t+1)} \right\|_F \tag{2}$$

where $\mathbf{R}^{(t)} \in \mathbb{R}^{d\times d}$ and doing $\mathbf{R}^{(t)}\mathbf{W}^{(t)}$ to adjust the embedding's axes on $t$ to those on $t+1$. Note that this preserves the cosine similarities between the columns of $\mathbf{W}^{(t)}$.

### Analysing semantic changes over time [3]

The rate of semantic change of $w_i$ at $t$ is defined as

$$\Delta^{(t)}\left(w_i\right) = \text{cos-dist}\left(\mathbf{w}_i^{(t)}, \mathbf{w}_i^{(t+1)}\right) \tag{3}$$

Then, we compute $\tilde{\Delta}^{(t)}(w_i)$, the normalized log-transformed rate of semantic change for a word $w_i \in \mathcal{V}$ at time $t \in \{t_0, \ldots, t_n\}$. This rate quantifies the semantic displacement $\tilde{\Delta}^{(t)}(w_i)$ occurring in a pair of consecutive decades, $t$ and $t+1$.

Finally, they fit a linear model to express $\tilde{\Delta}^{(t)}(w_i)$ is:

$$\begin{gathered}\tilde{\Delta}^{(t)}(w_i) = \beta_f \log\left(f^{(t)}(w_i)\right) + \beta_d \log\left(d^{(t)}(w_i)\right) \\ +\beta_t + z_{w_i} + \epsilon_{w_i}^{(t)}\end{gathered}$$

which assumes that word's frequency $f^{(t)}(w_i)$, polysemy $d^{(t)}(w_i)$, and decade $t$ impact the semantic change.

# 4 Our project

In a first moment, we'll fit coefficients $\beta_f, \beta_d, \beta_t$ using the standard maximum likelihood algorithm, following the authors' approach. We intend to verify that our results will reflect the two statistical laws of semantic evolution discovered by authors: *The law of conformity* and *The law of innovation*.

Next, using their pre-computed embeddings, we will propose a different linear model by including a new terms and/or replacing existing ones. For example, we consider initially focusing on the *relative frequency* and "*synonymyty*". The former will consist of transforming the word frequencies such that $\sum_{i=1}^{|\mathcal{V}|} f^{(t)}(w_i) = 1$. The latter would be some metric (to be defined) that measures the "amount of words close enough to $w_i$" - as if it captured the "inverse" of the polysemy.

Finally, consider the SGNS architecture as represented in Figure 1. We will add a second hidden layer to the model and compare both the shallow and deep embeddings with those of the original paper.

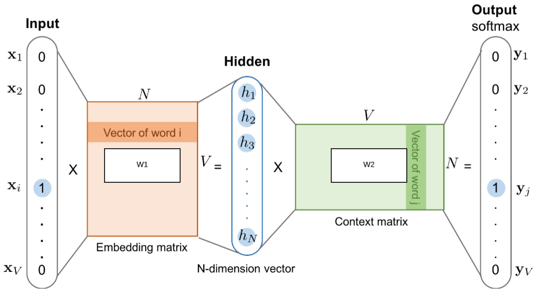

Figure 1: schematic diagram of the Word2Vec model (link to source)

For the sake of inspiration, we might try to reproduce some of what other authors have done in similar papers like [4], [8], [1], or [3].

[2] Here, $\mathcal{V}$ is the set of words (i.e. the vocabulary).

[3] This corresponds to the section 4 in the paper.

# References

[1] V. D. Carlo, F. Bianchi, and M. Palmonari. Training temporal word embeddings with a compass, 2019.

[2] M. Davies. The Corpus of Historical American English: 400 million words, 1810-2009. http://corpus.byu.edu/coha. 2010.

[3] J.-F. Delpech. Unsupervised detection of diachronic word sense evolution. 2018.

[4] H. Dubossarsky, D. Weinshall, and E. Grossman. Outta control: Laws of semantic change and inherent biases in word representation models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145, Copenhagen, Denmark, Sept. 2017. Association for Computational Linguistics.

[5] W. L. Hamilton, J. Leskovec, and D. Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, Aug. 2016. Association for Computational Linguistics.

[6] Y. Lin, J.-B. Michel, E. Aiden, J. Orwant, W. Brockman, and S. Petrov. Syntactic annotations for the google books ngram corpus. volume July, pages 169–174, 07 2012.

[7] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. 2013.

[8] Z. Yao, Y. Sun, W. Ding, N. Rao, and H. Xiong. Dynamic Word Embeddings for Evolving Semantic Discovery. page arXiv:1703.00607, Mar 2017.