W266 - Winter 2018

Time Aware Word Embeddings to Detect Language and Social Trends

Project Report

Vincenzo Moccia

In this project, we learn time aware word embeddings. Specifically, each word in a different time slice (e.g., years) is represented by a different vector. By looking at word neighborhoods in different time slices, we can observe word meaning and associations as they evolve over time and we can potentially infer language and social trends over different periods of human history.

Introduction

Language is constantly evolving, both across space and social groups, and across time. In particular, new words are borrowed or invented, and the meaning of old words drifts. Think about how the word *broadcast* changed its meaning after the radio and television were invented.

A number of large corpora are now available that extend across multiple years and carry precious information about language evolution. With modern language modeling techniques, we can process these corpora and obtain valuable insights. For example, changes in association of a certain word with positive or negative terms could track how people tend to feel about certain concepts (e.g., inventions like cars, political terms like socialism/capitalism, movements like civil-rights/environmentalism, religion etc.)

In this project, we learn word embeddings using the *word2vec* skip-gram model by Mikolov et al. [8] extended to the broader goal of time aware vectors. An important challenge is computing word embeddings for different time slices that can be compared meaningfully. As a regularization strategy to smooth embedding changes across time, we train embeddings for each year using those for the previous year as a starting point.

Finally, we find embeddings that show the most significant shift in time, and then cluster them to infer topics for the most relevant language and social trends in the analyzed time frame.

Background

There is a number of papers on the topic of time aware word embeddings, mostly proposing some time specific variation of the basic recurrent neural network language model described in Mikolov et al. [8].

Some researchers follow a two-step pattern: first compute static word embeddings in each time slice separately, then find a way to align the word embeddings across time slices (e.g. [1]). This approach suffers from the *alignment* problem, that makes trained embeddings from different time slices hard to compare.

Yao et al. [6], on the other hand, propose to learn temporal embeddings in all time slices concurrently and apply regularization terms to smooth embedding changes across time.

Similarly, Bamler and Mandt [2] propose a generalization of Mikolov's skip-gram model that they call the dynamic skip-gram model. This is a probabilistic model which combines a Bayesian version of the skip-gram model with a latent time series and finds word embedding vectors that continuously drift over time, allowing to track changes in language.

As an alternative approach to solving the alignment problem, we initialize and train our model on the first time slice, and then tune on subsequent time slices over multiple training epochs, as described below. This results in implicit regularization over time, since everything starts in the same place and then moves based on data for the following years.

Methods

This work is based on the New York Times Annotated Corpus available through the Linguistic Data Consortium at https://catalog.ldc.upenn.edu/ldc2008t19. This dataset contains over 1.8 million articles written and published by the New York Times between January 1, 1987 and June 19, 2007 (21 years). The text in this corpus is formatted in News Industry Text Format (NITF), an XML specification that provides a standardized representation for the content and structure of discrete news articles.

The corpus is a collection of documents organized by year, month and day. Each document includes rich metadata. For the purpose of this project, however, we only consider the body text of each article.

We use the data for year 1987 to train baseline embeddings. Then, we continue training with data for individual years, always starting from the embeddings for the previous year. To make sure the resulting embeddings are comparable, we train all models on the same number of epochs, so our model for year 1987 sees 1987 data 21 times; our model for year 1988 sees 1987 data once and then 1988 data 20 times, and so on.

More formally, let i be our time slices, with $i = 1, \ldots, n$, and $T(\text{input model, input data})$ our training function. We train multiple generations j of each model $M(i, j)$ using data $d(i)$ with $j = 1, \ldots, n-i+1$, where:

$$M(i, j) = \begin{cases} T(M(i, j-1), d(i)) & \text{if } j>1 \\ T(M(i-1, j-1, d(i)) & \text{if } j=1 \end{cases}$$

$M(0, 0)$ = model with random initialized parameters

Figure 1 illustrates the first training step in this process, where the first generation of our model for year 1987 is trained on 1987 data starting from random initialized parameters.

The second step is shown in figure 2. We build the second generation of our model for year 1987 by training the first generation model on 1987 data again. At the same time, we build the first generation of our model for year 1988 by training the first generation of model 1987 with 1988 data. Figure 3 shows the third step.

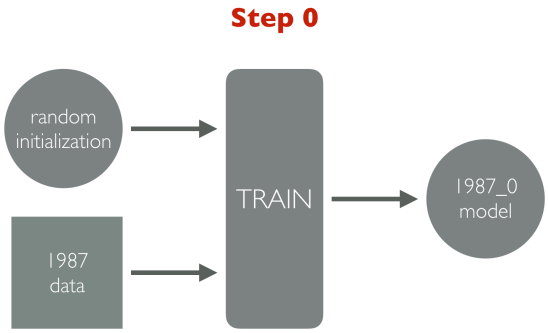

Figure 1. *The first generation of our model for year 1987 is trained on 1987 data starting from random initialized parameters.*

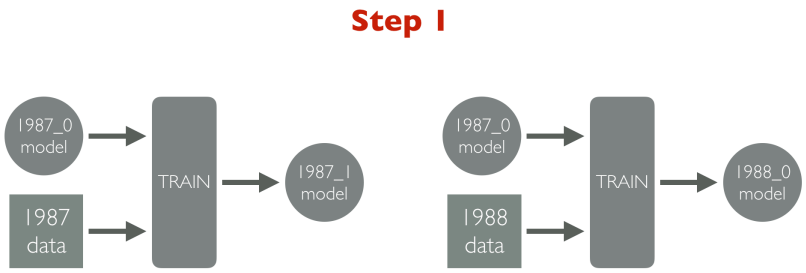

Figure 2. *We build the second generation of our model for year 1987 by training the first generation model on 1987 data again. Also, we build the first generation of our model for year 1988 by training the first generation of model 1987 with 1988 data.*

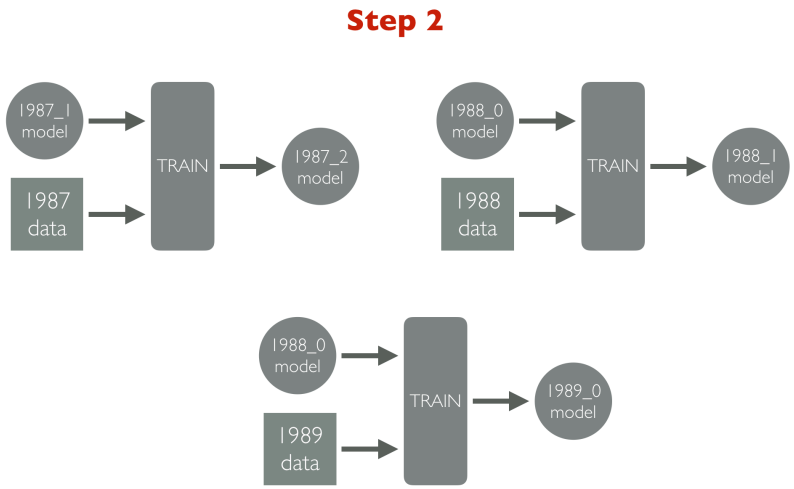

Figure 3. *The third generation of model 1987, second generation of model 1988 and first generation of model 1989.*

Our training function is based on the word2vec skip-gram model with negative sampling.

A key assumption is that interesting patterns can be detected even with word embeddings having a lower representation quality than would be required by more sophisticated applications, as long as temporal effects are preserved to some degree. Based on this assumption, we set our

embeddings to have size = 128. Also, we limit our vocabulary to the 50,000 most frequent words in the data set.

Results

Once embeddings for different time slices have been trained, we find words that show the most significant shift in the latent space over the time frame of 21 years. We use cosine distance as our metric. Since our model outputs normalized embeddings, we can compute cosine distance simply as the dot product between vectors.

$$\textit{index of word with biggest shift} = \text{argmax}_i \, \mathbf{w}_{i,1987} \, \mathbf{w}_{i,2007} \qquad i = 1, \ldots, \text{vocabulary size}$$

The top 5 words with the biggest shift are: *google, worldcom, hamas, www* and *qaeda*.

We take the top 500 words and then run an unsupervised clustering algorithm based on cosine distance (we use *AgglomerativeClustering* from *sklearn*). Each cluster represents an area for semantic shift that corresponds to a social or cultural trend in the corresponding time frame.

Three examples are presented below. We show visualizations for sample words created with a method similar to the one proposed in [1]: we collect a word's nearest neighbors in each time slice (1987 to 2007), then compute the t-SNE embedding of these words on the most recent time slice (2007).

Terrorism

This cluster includes words such as *hamas, qaeda, osama* and *islamist*. Our time frame includes the September 11 attacks and the profound impacts of this and other related events are clearly shown by our analysis.

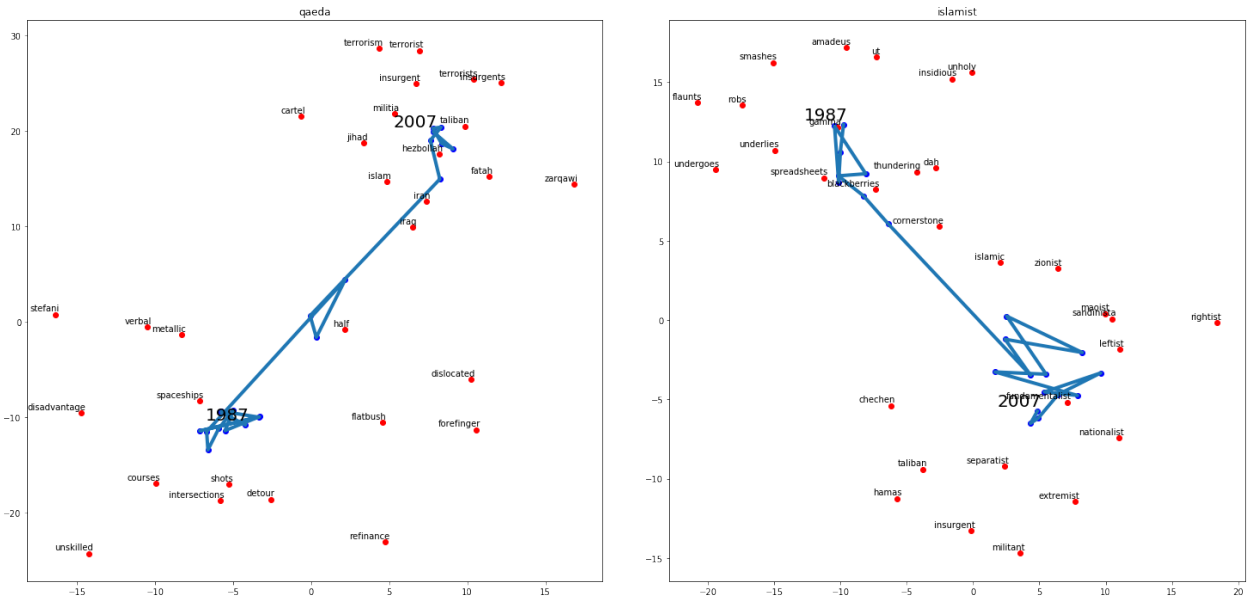

Figure 4. *Semantic shift of words "qaeda" and "islamist".*

Technology

Technology plays a fundamental role in the social and cultural trends of our time frame. As an example, the iPod was introduced in 2001. Some words in this cluster include *google, internet, cellphones* and *blog.*

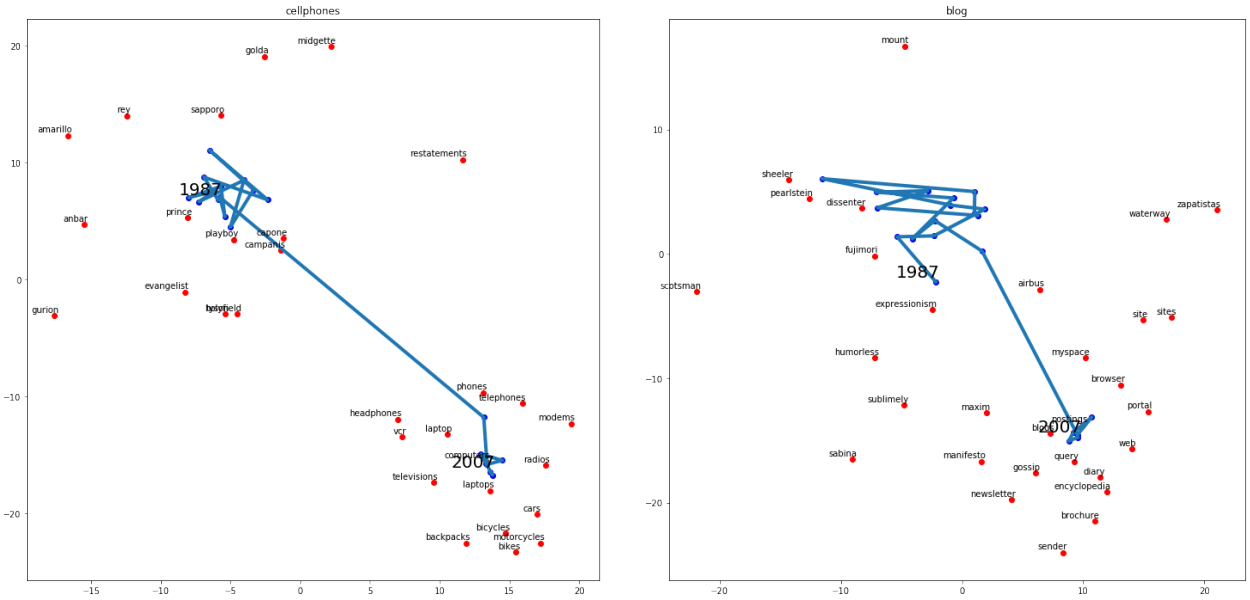

Figure 5. *Semantic shift of words “cellphones” and “blog”.*

Popular culture

This cluster includes words such as *buzz, vibe, mantra* and *flux.*

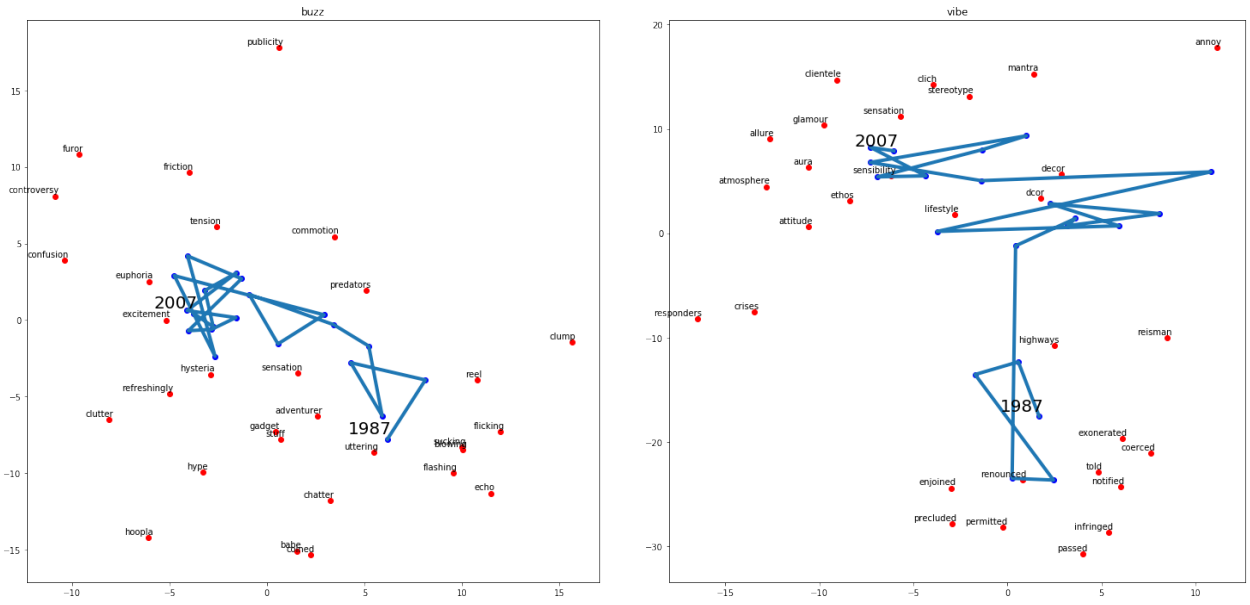

Figure 6. *Semantic shift of words “buzz” and “vibe”.*

Conclusion

In this paper we showed that vector representations of words derived through a time aware version of word2vec can be used to infer language and social trends. We proposed a simple and effective approach to creating such vectors that addresses the issue of alignment. Finally, we observed that it is possible to apply unsupervised clustering to vectors with the most significant shift in time to identify major areas for change in society and culture.

References

1. Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). **Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change**. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (pp. 1489–1501).
2. Bamler, R., & Mandt, S. (2017). **Dynamic Word Embeddings via Skip-Gram Filtering.** In Proceedings of ICML 2017. Retrieved from http://arxiv.org/abs/1702.08359
3. Dubossarsky, H., Grossman, E., & Weinshall, D. (2017). **Outta Control: Laws of Semantic Change and Inherent Biases in Word Representation Models.** In Conference on Empirical Methods in Natural Language Processing (pp. 1147–1156).
4. Szymanski, T. (2017). **Temporal Word Analogies : Identifying Lexical Replacement with Diachronic Word Embeddings.** In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (pp. 448–453).
5. Rosin, G., Radinsky, K., & Adar, E. (2017). **Learning Word Relatedness over Time.** In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.
6. 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 (WSDM '18). ACM, New York, NY, USA, 673-681.
7. Hosein Azarbonyad , Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, Jaap Kamps, **Words are Malleable: Computing Semantic Shifts in Political and Media Discourse**, Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, November 06-10, 2017, Singapore, Singapore
8. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013) **Efficient estimation of word representations in vector space**