TPPMI - a Temporal Positive Pointwise Mutual Information Embedding of Words

Paul Schmitt
TU Wien
paulschmitt.a@icloud.com

Zsófia Rakovics
Eötvös Loránd University
zsofia.rakovics@tatk.elte.hu

Márton Rakovics
Eötvös Loránd University
marton.rakovics@tatk.elte.hu

Gábor Recski
TU Wien
gabor.recski@tuwien.ac.at

Abstract

We present Temporal Positive Pointwise Mutual Information (TPPMI) embeddings as a robust and data-efficient alternative for modeling temporal semantic change. Based on the assumption that the semantics of the most frequent words in a corpus are relatively stable over time, our model represents words as vectors of their PPMI similarities with a predefined set of such context words. We evaluate our method on the temporal word analogy benchmark of Yao et al. (2018) and compare it to the TWEC model (Di Carlo et al., 2019), demonstrating the competitiveness of the approach. While the performance of TPPMI stays below that of the state-of-the-art TWEC model, it offers a higher degree of interpretability and is applicable in scenarios where only a limited amount of data is available.

1 Introduction

Word embedding models have become the dominant approach to modelling lexical semantics in the natural language processing (NLP) community. While contextual embeddings are now prevalent in most NLP applications, common static embedding methods such as word2vec (Mikolov et al., 2013) and GLoVe (Pennington et al., 2014) are still widely used in the computational modeling of word meaning, including the study of semantic change. Modern approaches train temporal word embeddings by learning alignments between multiple sets of word vectors (Hamilton et al., 2016; Di Carlo et al., 2019), but these approaches rely on the availability of a large amount of training data from each time period.

The efficiency and robustness of Pointwise Mutual Information (PMI) as a simple measure for word co-occurrence has been demonstrated in multiple studies (Bullinaria and Levy, 2007; Levy and Goldberg, 2014; Wendlandt et al., 2018). In this study we propose the use of Positive Pointwise Mutual Information (PPMI) to create temporal embeddings that represent the meaning of words as vectors of their PPMI with a small fixed set of context words chosen from the most frequent content words of the corpus, based on the assumption that the semantics of these words is relatively stable across time. Our experiments on the temporal word analogy task of Yao et al. (2018) demonstrate that this highly interpretable model offers a robust and competitive measure of lexical semantic change. The rest of the paper is structured as follows: Section 2 summarizes recent research on temporal word embeddings. Section 3 presents our method. Section 4 describes our experimental setup and Section 5 presents results of both quantitative and qualitative analysis. Section 6 concludes the paper. All software described here is publicly available on GitHub[1] under an MIT license.

2 Related Work

Word embeddings have been used extensively to study lexical semantic change. Yao et al. (2018) trains time-aware word embeddings by jointly learning multiple word embeddings and their alignment. For evaluation they train on a dataset of nearly 100,000 crawled articles from the New York Times (NYT), published between 1980 and 2016, and evaluate their method by using the resulting vector spaces to solve simple temporal reasoning tasks. One of these tasks that has since been reused for evaluating temporal embeddings, and which we also use in this paper, are *temporal analogy* questions of the form *2012:Obama = 2004:?*. In this example a temporal embedding is expected to predict *Bush* as a likely or even the most likely answer based on the assumption that the word's semantics in 2004 news texts should be (most) similar to that of *Obama* in 2012.

[1] https://github.com/FlackoJodye1/temporal-word-embeddings

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 119–125
July 5–10, 2020.

Rudolph and Blei (2018) develop Dynamic Bernoulli Embeddings, a type of Exponential Family Embeddings (Rudolph et al., 2016), which capture change by modeling words as sequences of embeddings over time slices that are grounded in a space of shared context vectors. They train their models on corpora of scientific papers and U.S. Senate speeches. In addition to qualitative analysis of the resulting embeddings they also perform intrinsic evaluation that involves calculating their loss function on heldout portions for each dataset and time period. This experiment is reproduced by Di Carlo et al. (2019), who propose the TWEC method for aligning word2vec embeddings trained on data from various time periods based on a shared target vector space trained on atemporal data. They also test their method on the temporal analogy task, and it is this approach that we use for comparison when evaluating the TPPMI method.

3 Method

The Temporal Positive Pointwise Mutual Information (TPPMI) method models semantic change of words based on their distribution w.r.t a fixed set of the most frequent content words of the atemporal context, based on the assumption that these words exhibit relatively stable semantics. Pointwise Mutual Information (PMI) measures the co-occurrence of a word w with a context word c by calculating

$$\text{PMI}(w, c) = \log \frac{\hat{P}(w, c)}{\hat{P}(w)\hat{P}_\alpha(c)} - \log(k)$$

where $\hat{P}(w, c)$ is the co-occurrence probability of w and c, $\hat{P}(w)$ is the overall probability of w, $\hat{P}_\alpha(c)$ is the probability of c smoothed and k is a shifting constant (Levy and Goldberg, 2014). In all our experiments we use $\alpha = 1$ and $k = 1$. Positive Pointwise Mutual Information (PPMI) is defined as

$$\text{PPMI}(w, c) = \max(\text{PMI}(w, c), 0)$$

TPPMI embeddings for each time period map words to vectors of PPMI values between each word and the fixed set of context words, calculated on data from the given time period. This results in word embeddings that are highly interpretable compared to standard word vectors, since dimensions directly correspond to individual context words. As a second step, the entries of the PPMI matrices are smoothed in time using a cubic spline separately for each component of the embedding vectors to stabilize the vectors in each slice.

The static set of context words is determined by removing stopwords from the atemporal corpus (the union of all time slices) and sampling from the most frequent words in the corpus. The number of words, which determines the dimensionality of the TPPMI embeddings, is a parameter of our approach. To create a set of n context words we sample from the $2n$ most frequent words. Stopword removal is performed using the `nltk`[2] package. The size of the context word set greatly influences the robustness and performance of our models and should be optimized separately for each application of the TPPMI approach.

4 Experiment

Following the experimental setup of Di Carlo et al. (2019) we train our temporal embeddings on the NYT dataset and evaluate it on temporal word analogies (see also Section 2). We compare our model to both TWEC and to static word2vec embeddings as a trivial baseline.

4.1 Models

The TPPMI embeddings are trained using the process described in Section 3. The number of context words is set to 2,000. The TWEC model and the static word2vec model (SW2V) are trained usting the hyperparameters from Yao et al. (2018) and Di Carlo et al. (2019), embedding dimension is 50, the context window size is 5, and the vocabulary size is 21,000. All text is lowercased, stopwords as well as words with an overall frequency below 200 are omitted.

4.2 Temporal Word Analogies

We compare the TPPMI model with established methods using a modified version of the temporal analogical reasoning task introduced by Yao et al. (2018). The task of solving a temporal word analogy (TWA) can be expressed as $t1 : w1 = t2 :?$ and entails predicting the word $w2$ that at time $t2$ is semantically most semantically similar to the word $w1$ at time $t1$. In all vector space models this prediction is achieved by identifying the word whose vector in the vector space of time $t2$ is most similar to the vector of $w1$ at time $t1$.

[2] `https://www.nltk.org/`

The training dataset contains 99,872 crawled articles from the New York Times, all of them published between January 1990 and July 2016. The dataset was also used by Yao et al. (2018) and was provided to us by the authors. Following previous experiments we partition the articles into batches for each calendar year, resulting in a total of 27 slices. The temporal analogy queries introduced by Yao et al. (2018) are derived from publicly available records and contain the names of persons occupying various public offices in each calendar year, including U.S. President, the Chancellor of Germany, the Governor of New York, among others. In our experiments we focus only on analogies involving U.S. Presidents. The test queries contain two types of analogies:

- **Static analogies:** The target word is identical to the query word, e.g. *2003:bush = 2004:bush*
- **Dynamic analogies:** The target word differs from the query word, e.g. *2003:bush = 2011:obama*

Following Di Carlo et al. (2019) we evaluate our method separately on each subset. This is necessary to separate cases where the trivial strategy of the static embedding (SW2V) yields the correct answer. Evaluating on both datasets ensures that temporal embedding models strike a balance between stability and dynamism. Basic descriptive statistics about the test set are shown in Table 1.

Analogies	Total queries	Unique queries
All	8272	369
Static	2333	335
Dynamic	5938	369

Table 1: Basic statistics of the Temporal Word Analogy test set. For each unique pair of query word and year (e.g. *2012:obama*) the test set contains queries for multiple years (e.g. *1990:?*, *2000:?*, etc.), hence the total number of queries is much larger then the number of unique queries

Named Entities Our early experiments revealed a significant artefact of the evaluation data. Since all queries and target words are named entities, evaluation results are largely influenced by some models' tendency to predict target words that have the same part-of-speech as the query word, behavior that is characteristic of most static word embeddings. Since this behavior offers an unwanted advantage on the TWA task, we modify the experimental setup by filtering words predicted by any model to only contain named entities. This strategy increases the performance of all models, since the set of possible answers is considerably reduced, but focuses the evaluation on models' ability to predict semantic shifts. For the filtering step we use the Pantheon dataset of globally famous biographies (Yu et al., 2016), the set of possible target words is reduced to those that are listed in this dataset as person names. This strategy can trivially be extended to other entity types to allow for broader sets of TWA queries.

4.3 Evaluation

For each model cosine similarity is used to retrieve the vectors most similar to that of the target word, yielding a ranked list of possible answers to each query. These lists are then compared to the ground truth using two metrics, Mean Reciprocal Rank over the top 10 answers (MRR@10) and Mean Precision at various thresholds (MP@k). Both metrics are defined below.

Mean Reciprocal Rank (MRR@10) is the average rank that a model assigns to the correct answer. For each query i, rank_i is the rank of the expected answer in the list of predicted answers returned by a model. The MRR of the model can then be defined as

$$\text{MRR} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_i}$$

To calculate MRR@10, $\frac{1}{\text{rank}_i}$ is set to 0 if the word is not among the top 10 predicted words.

Mean Precision (MP@k) averages over all queries whether the expected answer is among the top k predicted answers. For a query i we define $P_i@k$ to be 1 if the top k predicted words contain the target word and 0 otherwise. MP@k is then defined as

$$\text{MP@}k = \frac{1}{N}\sum_{i=1}^{N} P_i@k$$

MP@1 is equivalent to model accuracy, measuring the ratio of queries for which the model successfully predicted the target word as the most likely answer.

5 Results

5.1 Quantitative Analysis

Table 2 shows all scores for each of the three models. The static baseline (SW2V) that uses the same vector space for query and target words achieves 1.0 accuracy (MP@1) on the static test set and 0.0 accuracy on the dynamic test set. Its MP@3 score of 0.709 on the dynamic test set demonstrates that most target words are among those that are distributionally most similar to the query word in the atemporal space, i.e. all names of recent U.S. presidents are relatively close together in a static word embedding. This property of the corpus together with our NE filtering strategy is responsible for high MP@k scores across the board, MP@10 values show that for all models the target word is among the top 10 predicted words for 80% of static queries and between 48 and 56% of dynamic queries. The high scores achieved even by the trivial baseline SW2V on the complete dataset ("All") also illustrates the need to evaluate models separately on the dynamic subset, i.e. on those analogies where the target word is different from the query word.

Both the TWEC and TPPMI models perform robustly in static analogies, with TWEC achieving slightly higher scores. TWEC's MRR@10 is 0.668 compared to TPPMI's 0.592, and TWEC's MP@1 is 0.591 compared to TPPMI's 0.493. However, TPPMI shows strong performance with an MP@3 of 0.663 and MP@5 of 0.729, demonstrating its capability to rank relevant words highly in static contexts. This indicates that while both models effectively capture stable semantic associations, TWEC has a slight edge in precision. Nonetheless, the TPPMI model showcases its ability to produce robust temporal embeddings with a much simpler approach.

On dynamic analogies, the TWEC model significantly outperforms TPPMI, achieving an MRR@10 of 0.402 and MP@1 of 0.326 compared to TPPMI's 0.302 and 0.225, respectively. In terms of MRR the TPPMI is on par only with the static baseline, but its accuracy (MP@1) of 0.225 on the dynamic set indicates its potential for correctly predicting semantic shifts. While further research shall be necessary to improve our method, these preliminary results suggest that the TPPMI model has potential as a simple, interpretable, and computationally efficient alternative to state-of-the-art methods. The interpretability of the method is further demonstrated by the qualitative analysis in the next section.

5.2 Qualitative analysis

Much recent work on temporal word embeddings has performed qualitative analysis using a variety of trajectory visualizations based on 2-dimensional projections of vector spaces. In our work we focus only on relative similarity of vectors as measured by cosine similarity and conduct two simple experiments for inspecting our model's ability to capture semantic change and temporal analogies, respectively.

Figure 1 plots the cosine similarity between the word "president" and the names "obama," "biden," "clinton," and "bush" over the years 1990 to 2016. The gray dotted lines on the graph indicate the years when a new president was elected: Bill Clinton in 1992, George W. Bush in 2000, and Barack Obama in 2008. This plot is especially interesting because Bush is also the name of the U.S. President before 1992 and Clinton is also the name of the Democratic candidate in 2016, accounting for the periodicity observed in each curve.

Next we demonstrate the workings of a temporal word analogy. Given the TWA query *2004:Bush = 2012:?* the prediction of the TPPMI model will be based on the similarity of target words in 2012 to those context words that are most similar to *Bush* in 2004. Figure 2 shows the top 10 such context words and their similarities to both *Bush* in the 2004 vector space and to *Obama* in the 2015 vector space. The years 2004 and 2012 were chosen as they are the re-election years for George W. Bush and Barack Obama, respectively. We can observe that some, but not all of these context words maintain a high similarity with the name of the sitting president across time periods. While in this case the observed distinctions are trivial, e.g. that among the words most closely associated with *Bush*, *president* and *re-election* are more distinctive of his 2012 role than the word *George*, it nevertheless demonstrates the TPPMI model's ability to offer similar but less trivial insights from limited amount of temporal data.

6 Conclusion

We presented the Temporal Positive Pointwise Mutual Information model of lexical semantic change. TPPMI offers an interpretable and robust approach to capturing temporal semantic shifts of words, addressing the challenges of small and sparse datasets.

Table 2: Evaluation results on the Temporal Word Analogy task.

Model	Category	MRR@10	MP@1	MP@3	MP@5	MP@10
TWEC	Static	0.668	0.591	0.723	0.768	0.818
	Dynamic	**0.402**	**0.326**	**0.455**	**0.508**	**0.560**
	All	0.455	**0.383**	0.504	0.551	0.602
TPPMI	Static	0.592	0.493	0.663	0.729	0.791
	Dynamic	0.302	0.225	0.348	0.409	0.475
	All	0.365	0.284	0.417	0.478	0.541
SW2V	Static	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
	Dynamic	0.322	0.000	0.709	0.741	0.813
	All	**0.551**	0.337	**0.807**	**0.828**	**0.876**

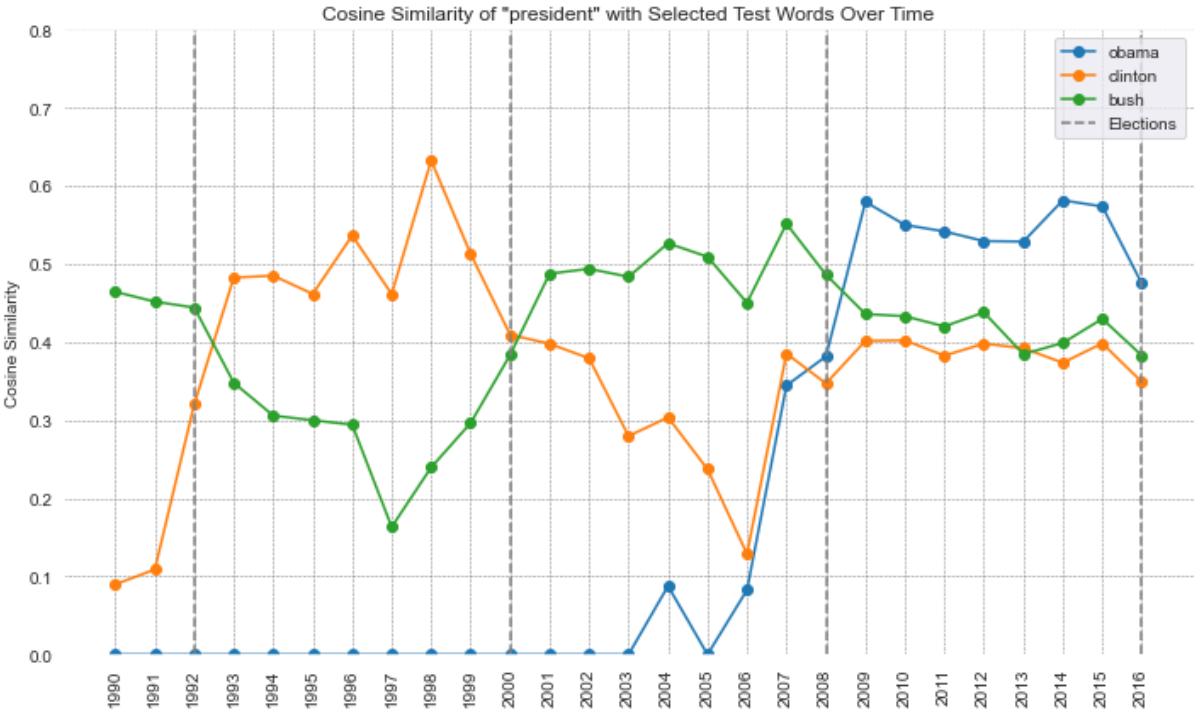

Figure 1: Yearly cosine similarities between the word 'president' and the names of U.S. Presidents between 1990 and 2016, as measured by the TPPMI model

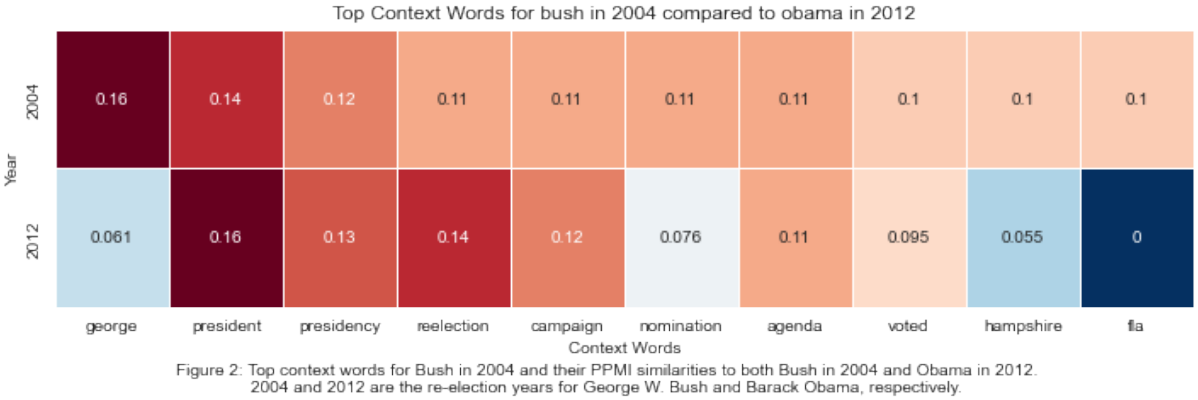

Figure 2: Top context words for Bush in 2004 and their PPMI similarities to both Bush in 2004 and Obama in 2012. 2004 and 2012 are the re-election years for George W. Bush and Barack Obama, respectively.

Figure 2: Top context words for *Bush* in 2004 and their PPMI similarities to both *Bush* in 2004 and *Obama* in 2012. 2004 and 2012 are the re-election years for George W. Bush and Barack Obama, respectively.

The model is evaluated on a temporal word analogy task and achieves reasonable performance on both static and dynamic analogies. Despite its inferiority to more sophisticated models like TWEC, we believe that its simplicity and computational efficiency make TPPMI a practical alternative for applications with limited data. Our qualitative analysis further demonstrates the model's ability to show semantic shifts of individual words over time and to offer explanations of such shifts based on the words corresponding to significant dimensions.

Limitations

Despite its strengths, the TPPMI model's performance is clearly limited and appears to be inferior to state-of-the-art methods on the TWA benchmark. While the method is a practical alternative for applications with limited data and a need for explainability, it is likely not sufficiently robust for large-scale analysis of semantic change. The significance of this preliminary work is further limited by the choice of a single training dataset, a single evaluation benchmark, and a single reference system.

Ethical considerations

As any distributional model, TPPMI embeddings may inherit and amplify harmful biases present in its training data. Mitigating this risk requires careful data selection, preprocessing, and ongoing evaluation of model bias. However, the interpretability of TPPMI embeddings offers a lowered risk of bias in temporal predictions compared to alternative methods, since the significant dimensions are directly associated with individual context words.

Notes

The first version of the TPPMI method was presented at the conference of ELTE Angelusz Róbert College for Advanced Studies in Social Sciences and published in the associated conference proceedings (Rakovics, 2022). The improved version of the method was presented at the 8th International Conference on Computational Social Science (IC2S2) as a conference poster (Rakovics and Rakovics, 2022).

References

John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526.

Valerio Di Carlo, Federico Bianchi, and Matteo Palmonari. 2019. Training Temporal Word Embeddings with a Compass. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6326–6334.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. 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. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Zsófia Rakovics. 2022. A Temporal Positive Pointwise Mutual Information (TPPMI) időbeli szóbeágyazási modell alkalmazásában rejlő lehetőségek demonstrálása. In *Van új a nap alatt: Az ELTE Angelusz Róbert Társadalomtudományi Szakkollégium konferenciájának tanulmánykötete*, pages 31–48. ELTE Eötvös József Kiadó; ELTE Angelusz Róbert Társadalomtudományi Szakkollégium.

Zsófia Rakovics and Márton Rakovics. 2022. Semantic evolution of words in Hungarian Prime Minister Viktor Orbán's speeches using a temporal word embedding model focusing on the issue of migration.

Maja Rudolph and David Blei. 2018. Dynamic Embeddings for Language Evolution. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, pages 1003–1011, Lyon, France. ACM Press.

Maja Rudolph, Francisco Ruiz, Stephan Mandt, and David Blei. 2016. Exponential family embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Laura Wendlandt, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. Factors influencing the surprising instability of word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2092–2102, New Orleans, Louisiana. Association for Computational Linguistics.

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, page 673–681, New York, NY, USA. Association for Computing Machinery.

Amy Zhao Yu, Shahar Ronen, Kevin Hu, Tiffany Lu, and Cesar Hidalgo. 2016. pantheon.tsv. In *Pantheon 1.0, A Manually Verified Dataset of Globally Famous Biographies*. Harvard Dataverse.