CARMA 2025

7th Int. Conf. on Advanced Research Methods and Analytics
2-4 July 2025, Rome, Italy
DOI: https://doi.org/10.4995/CARMA2025.2025.20553

Time Series Analysis of Temporal Word Embeddings: Geometric vs Probabilistic Interpretations

Nauel Serraino, Riccardo Ricciardi, Paola Zuccolotto
Department of Economics and Management, University of Brescia, Italy.

How to cite: Serraino, N.; Ricciardi, R.; Zuccolotto, P. 2025. Time Series Analysis of Temporal Word Embeddings: Geometric vs Probabilistic Interpretations. In: 7th International Conference on Advanced Research Methods and Analytics (CARMA 2025). Rome, 2-4 July 2025. https://doi.org/10.4995/CARMA2025.2025.20553

Abstract

This study explores the probabilistic interpretation of temporal word embeddings to investigate semantic evolution. Traditional embeddings rely on words as fixed vectors, limiting their ability to capture linguistic variability over time. By presenting embeddings as probability distributions, we can better evaluate the uncertainty of word meanings. Using temporal embeddings and applying entropy-based measures like Shannon entropy and Kullback-Leibler divergence, we quantify semantic shifts across time. This probabilistic interpretation, combined with the creation of Fixed Base and Chain Base time-series derived from index number, provides a new framework for analyzing semantic evolution.

Keywords: *Temporal Word Embeddings; Semantic Evolution; Fuzzy Entropy; Time-series.*

1. Introduction

The probabilistic interpretation of temporal word embeddings offers a novel and powerful approach to studying semantic evolution. The traditional methods of embedding words in vector spaces usually treat them geometrically. However, by interpreting word embeddings as probability distributions, we can embrace the fuzziness of word meanings and better reflect the evolving nature of language. This probabilistic perspective allows us to determine the likelihood of a word belonging to certain semantic features, providing a better understanding of how meanings change across time.

Recent research has demonstrated the benefits of embedding words as probabilistic distributions, particularly through the normalization of embedding matrices. This approach, as outlined in the work of (Bhat et al. 2020), treats embeddings as fuzzy sets of Bernoulli random variables, enabling the use of entropy-based measures, such as Shannon entropy and Kullback-Leibler (KL) divergence, to quantify uncertainty and divergence in word meanings over time.

These measures can be used to create univariate time-series, as pointed out by (Aghabozorgi et al., 2015). In fact, a simple approach to calculating the distance between two time-series is to treat them as univariate series and compute the distance measure at each time point.

This work draws different conclusions. Firstly, after a comparison between the two ways of interpreting temporal embeddings, it highlights the potential benefits of the probabilistic interpretation in the study of the evolution of temporal word embeddings. Secondly, it incorporates the study of the evolution of word embeddings within the framework of univariate time series. In this way, similarities over time can be treated as index numbers to analyze both short-term and long-term semantic changes in words.

Our framework can be applied across various domains of knowledge. We propose it in a generic context of the temporal evolution of different corpora. Nonetheless, the same principles can be applied to diverse scenarios, such as changes in advertising communication, shifts in political messaging, or variations in product reviews and consumer feedback.

2. Related Work

In the rapidly evolving field of natural language processing, understanding how language changes over time have become increasingly important. Traditional word embeddings, which are effective in capturing semantic relationships between words, struggle to account for the shifts in meaning that occur as language evolves. In fact, as pointed out by (Yao et al. 2018), the use of independently trained word2vec-like models on corpora belonging to different time periods usually suffer from the alignment problem, in which the different embeddings do not align on the same latent space, making the comparison of different word representations impossible.

To address this limitation, we rely on the Temporal Word Embedding with a Compass (TWEC) method proposed by (Di Carlo et al. 2019) to train temporal word embeddings. It is a modification of the well-known Continuous Bag-of-Words (CBOW) variant of Word2Vec proposed by (Mikolov et al. 2013). This architecture is based on the following heuristic: training a word embedding on a general corpus, then using this embedding to stabilize one of the layers of the CBOW architecture used to train the subsequent temporal embeddings. The resulting embeddings aim to represent words in a way that reflects the changes in meanings across different time periods, overcoming the alignment issue of the distinct matrices.

3. Methods and data

To analyze how the meaning of words changes over time, we can model their semantic evolution. Suppose the corpus $\mathcal{D}$ is divided into T subcorpora $\mathcal{D}_t$, where $t = 1, \dots, T$ represents different time periods. To study the semantic evolution of words over time, a temporal embedding function maps words into a time-varying $\mathbb{R}^d$ semantic space. In this case, each word w_i is represented by a vector $e^t(w_i)$, belonging to a matrix W^t of temporal word embeddings for each time step t:

$$W^t = \begin{bmatrix} e^t(w_1) \\ e^t(w_2) \\ \dots \\ e^t(w_v) \end{bmatrix} \in \mathbb{R}^d \tag{1}$$

A temporal embedding function maps words that appear in similar contexts at a given time step to similar vector representations, and simultaneously, for two different time points, assigns similar representations to a word that appears in similar contexts at both times. As a result, the similarity relationships among word vectors, both between time periods and within the same period, are used to examine how the semantics of words and the semantic relationships between words change over time.

A standard measure of pairwise similarity between vectors is the cosine of the angle formed by two vectors that is widely used in the literature, like in (Levy et al. 2014). In general, we can compute the cosine similarity between two generic words embeddings $e^t(w_i)$, and $e^t(w_j)$[1], extracted from the temporal embedding matrix W^t, using the following formula:

$$\text{cosine}\left(e^t(w_i), e^t(w_j)\right) = \frac{e^t(w_i) \cdot e^t(w_j)}{|e^t(w_i)||e^t(w_j)|} \tag{2}$$

[1] The same formulation can be applied to the word embedding representation at different time t.

that is the dot product between the vectors divided by the product of their lengths, which ranges in [−1,1].

However, in recent advancements in embedding techniques, such as those introduced by (Bhat et al. 2020), demonstrate innovative approaches to representing words as probability distributions. These innovation offers a more dynamic representations of language by considering the broader linguistic context in which words appear.

Probabilistic interpretation takes place through the normalization, column-wise, of the embedding matrix W^t. In such a way, each row vector $e^t(w_i)$ can be viewed as a fuzzy set of Bernoulli random variables, in the sense defined by (L.A. Zadeh, 1968), where each of the embedding feature has a probability p_i to be associated to the ith semantic feature. This interpretation enables the adoption of the concept of fuzzy entropy and the use of similarity measure like the KL divergence.

Firstly, the entropy of a single word A represented by a fuzzy set of Bernoulli random variables can be formulated as follows:

$$H(A) \stackrel{\text{def}}{=} \sum_i -p_i \ln p_i - (1-p_i)\ln(1-p_i) \tag{3}$$

In our scenario, entropy effectively captures language features like polysemy: the higher the entropy, the greater the variation in semantic context. In this way, entropy can effectively signal words that might be particularly suitable for further analysis.

While the KL divergence between two distributions S and T, representing two distinct words' fuzzy sets of Bernoulli random variables, can be formulated as:

$$D(S \parallel T) \stackrel{\text{def}}{=} \sum_i \left[p_i^S \ln\left(\frac{p_i^S}{p_i^T}\right) + (1-p_i^S)\ln\left(\frac{1-p_i^S}{1-p_i^T}\right)\right] \tag{4}$$

Specifically, higher values in KL divergence indicate greater change in semantic shift, while lower values suggest stability.

We will investigate the behavior of single words over time by computing cosine similarity and KL divergence between their vectors. Interpreted as index numbers on two different bases, these are standard tools in economics for measuring aggregate change (Balk, 2008).

Table 1. Different types of time-series applied.

Name	Formulation	Description
Fixed Base	$M\left(e^t(w_i); e^0(w_i)\right)$	*Measures the change relative to the baseline* $t = 0$.
Chain Base	$M\left(e^t(w_i); e^{t-1}(w_i)\right)$	*Measures the change relative to the previous period* $t - 1$.

By leveraging these index numbers, we can create different types of univariate time-series, from which we can identify multiple explanatory patterns in the temporal evolution of word meanings: while a Fixed-Base (FB) index number allows us to capture long-term semantic changes, a Chain-Base (CB) index number highlights short-term variations.

The underlying dataset for this study originates from a standardized corpus provided by The New York Times2, made up of UTF-8 encoded newspaper titles and excerpts from articles published between 1920 and 2020. The entire dataset, before the phase of text pre-processing, has a dimension of approximately ~2.5GB. After preprocessing, which included tokenization, lemmatization, and filtering out words with fewer than two letters, the resulting dataset provides a rich representation of the evolving language.

4. Results

The preliminary results, which compare the distribution of cosine similarity and KL divergence over time by means of boxplots and are presented in Figures 1-2, demonstrate that under both cosine similarity and KL divergence metrics, the vector representations of words become increasingly misaligned over time for the FB time-series. This is not surprising: as language evolves, the meanings and contextual usage of words shift, leading to greater divergence from

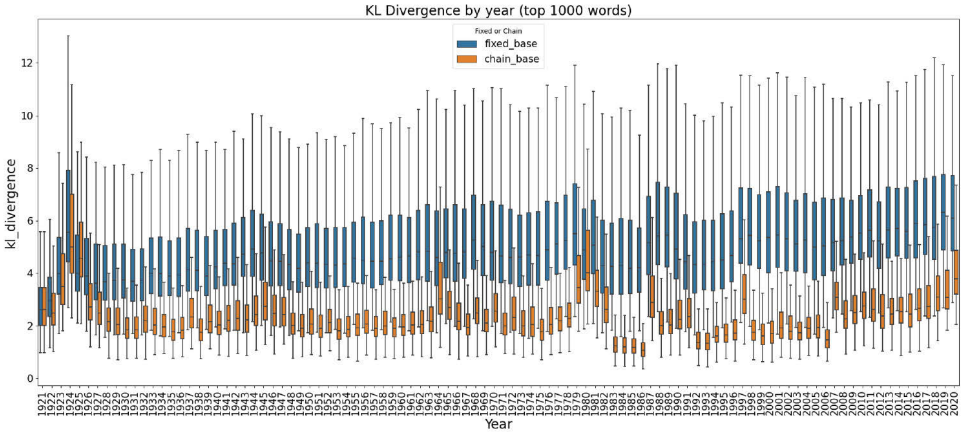

Figure 1. KL Divergence by Year: word embedding with itself over time, considering the first 1000 words per each year.

their original embeddings. Instead, the CB time series shows a pattern of small, steady changes occasionally interrupted by sudden spikes, which may suggest a shift in editorial style. Moreover, while both metrics show a decreasing alignment trend, the effect is more pronounced with cosine similarity, where semantic drift leads to greater divergence from the FB representation.

[2] Dataset available here: https://www.kaggle.com/datasets/tumanovalexander/nyt-articles-data

This suggests that cosine similarity and KL divergence highlight different aspects of semantic change.

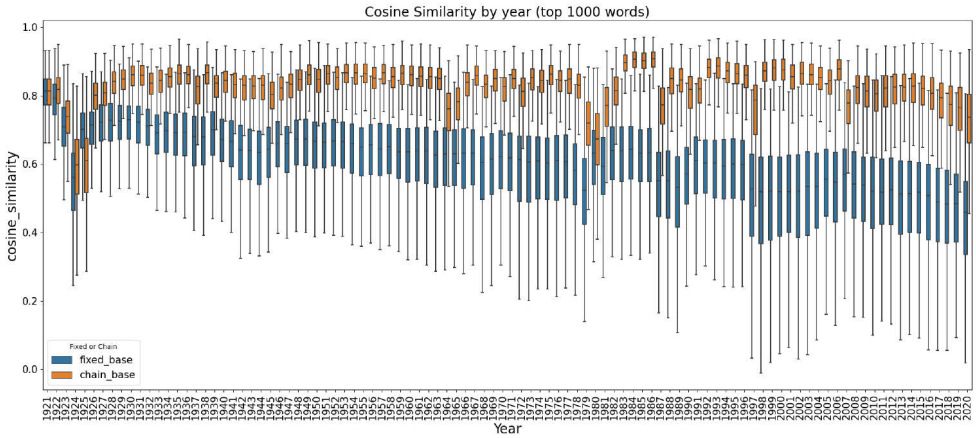

Figure 2. Cosine Similarity by Year: word embedding with itself over time, considering the first 1000 words per each year.

Further analysis explored the potential of probabilistic interpretation as a tool for detecting sudden shifts in the semantic meaning of words. Specifically, we examined the behavior of the two metrics applied to a list of United States presidents over the observed time span as shown in Figure 3.

KL divergence reveals distinct spike patterns during presidential mandates, typically showing low variance in semantic shift both before and after their tenure. However, some politicians exhibit continuous semantic shifts even beyond their presidency, potentially indicating a lasting political legacy, where public debate remains focused on certain figures (e.g., Roosevelt, Kennedy, Nixon), or cases of homonymous names (e.g., Bush, Clinton). In contrast, cosine similarity struggles to clearly identify these shifts, instead displaying generally higher variance throughout the analyzed period.

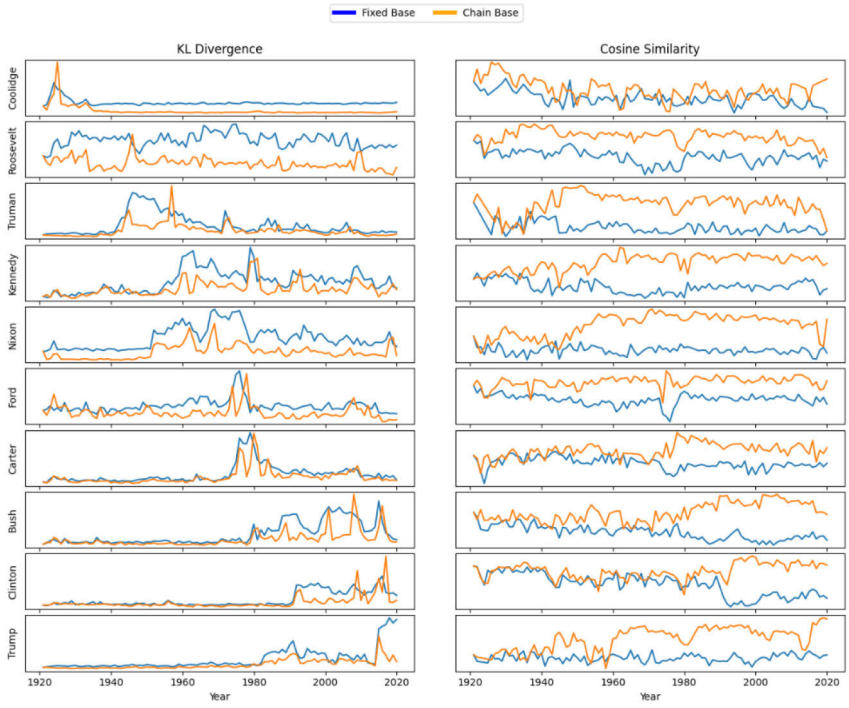

Figure 3 - Shift in the vectorial representation of United States presidents' names over time, measured using both KL divergence and Cosine similarity.

5. Conclusions and Discussion

This paper introduces a framework for analyzing the evolution of word embeddings over time using probabilistic methods and univariate time-series analysis. This approach can be applied to fields like sentiment analysis, product reviews, and political messaging to better understand linguistic changes.

Our preliminary results show that word meanings evolve significantly over time and that probabilistic interpretation can provide a richer representation of such shifts. To validate our findings, we aim to apply this interpretation to ground truth evaluations to identify the contexts in which this innovation is most effective.

Future work can extend this framework by incorporating additional metrics, such as the Jaccard Score, considering both its geometric and probabilistic interpretation. Furthermore, we plan to apply time-series clustering techniques to these metrics to distinguish between words that remain stable over time, those that exhibit short-term fluctuations, and those that undergo long-term semantic shifts.

References

Aghabozorgi, S., Seyed Shirkhorshidi, A., & Ying Wah, T. (2015). Time-series clustering – A decade review. Information Systems, 53, 16–38. https://doi.org/10.1016/j.is.2015.04.007

Balk, Bert. (2008). Price and quantity index numbers. Models for measuring aggregate change and difference. Reprint of the 2008 hardback ed. 10.1017/CBO9780511720758.

Bhat, S., Debnath, A., Banerjee, S., & Shrivastava, M. (2020). Word Embeddings as Tuples of Feature Probabilities. Proceedings of the 5th Workshop on Representation Learning for NLP, 24–33. https://doi.org/10.18653/v1/2020.repl4nlp-1.4.

Di Carlo, V., Bianchi, F., & Palmonari, M. (2019). *Training Temporal Word Embeddings with a Compass* (No. arXiv:1906.02376). arXiv. https://doi.org/10.48550/arXiv.1906.02376.

Levy, O., & Goldberg, Y. (2014). Linguistic Regularities in Sparse and Explicit Word Representations. In R. Morante & S. W. Yih (Eds.), *Proceedings of the Eighteenth Conference on Computational Natural Language Learning* (pp. 171–180). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-1618.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (No. arXiv:1301.3781). arXiv. https://doi.org/10.48550/arXiv.1301.3781.

Yao, Z., Sun, Y., Ding, W., Rao, N., & Xiong, H. (2018). Dynamic Word Embeddings for Evolving Semantic Discovery. *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 673–681. https://doi.org/10.1145/3159652.3159703.

Zadeh, L. A. (1968). Probability measures of Fuzzy events. *Journal of Mathematical Analysis and Applications*, *23*(2), 421–427. https://doi.org/10.1016/0022-247X(68)90078-4.