Exploring Word Usage Change with Continuously Evolving Embeddings

Franziska Horn
Machine Learning Group, Technische Universität Berlin, Berlin, Germany
BIFOLD, Berlin Institute for the Foundations of Learning and Data, Berlin, Germany
franziska.horn@alumni.tu-berlin.de

Abstract

The usage of individual words can change over time, for example, when words experience a semantic shift. As text datasets generally comprise documents that were collected over a longer period of time, examining word usage changes in a corpus can often reveal interesting patterns. In this paper, we introduce a simple and intuitive way to track word usage changes via continuously evolving embeddings, computed as a weighted running average of transformer-based contextualized embeddings. We demonstrate our approach on a corpus of recent New York Times article snippets and provide code for an easy to use web app to conveniently explore semantic shifts with interactive plots.

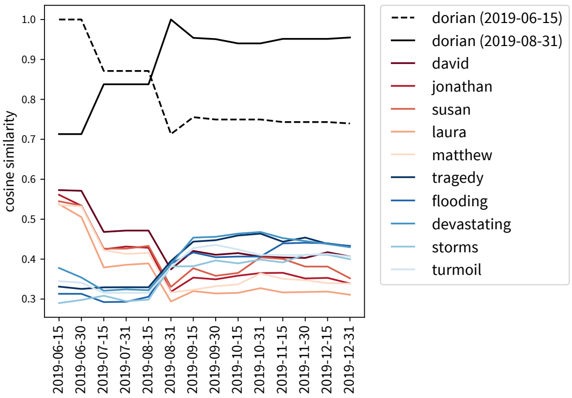


Figure 1: Cosine similarity between the continuously evolving embeddings for the word “Dorian” and its nearest neighbors over time, computed on our NYTimes article snippets corpus (see Sec. 4 for further details).

1 Introduction

Languages are constantly changing, with new words being coined or existing ones adopting a new meaning (Blank, 1999; Hamilton et al., 2016). For example, as Hurricane Dorian hit the Bahamas on Sept. 1, 2019, and was henceforth regarded as the worst natural disaster in the country’s recorded history, within a matter of days the until then innocuous name “Dorian” suddenly became synonymous with a devastating tropical cyclone (Fig. 1). These kinds of semantic shifts are of great interest for researchers in fields like computational linguists and digital humanities, but their analysis requires appropriate tools, especially to create fine-granular visualizations, for example, to facilitate the study of texts from fast-paced environments such as social media.

Word embeddings are nowadays the method of choice when examining the meaning of and relation between words, and, as an extension thereof, diachronic embeddings can be used to discover and analyze the semantic shifts and usage changes of words over time. The main idea behind diachronic word embeddings is to learn a set of embeddings for each word, one for each time period of interest, to then see how much the embeddings for the same word differ over time (see e.g. (Kutuzov et al., 2018) or (Tahmasebi et al., 2018) for a comprehensive overview). However, most approaches for computing diachronic embeddings either a) rely on static word embedding models such as word2vec, which makes it difficult to use them with small corpora, b) are based upon rather complex dynamic language models, and/or c) require the corpus to be split into individual time slices, which introduces a bias, since by computing embeddings for different years, for example, one implicitly assumes that the meaning of a word might change between January and December of the previous year, but not between July and August of the same year.

In this paper, we introduce continuously evolving embeddings that are computed in one pass over the whole (chronologically ordered) corpus by keeping track of a weighted running average of contextualized embeddings generated by a transformer model such as BERT (Sec. 2). By taking (poten-

Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations, pages 290–297, August 1st - August 6th, 2021.

tially arbitrarily frequent) ‘snapshots’ of the current state of the embeddings at user-defined time points, one obtains smoothly changing high-resolution diachronic embeddings. With these embeddings, semantic shifts can be detected at a resolution of weeks or months instead of years or decades. The exploration of word usage change is facilitated by our web app that provides the user with the corresponding interactive graphics (Sec. 3), which we demonstrate on a corpus of recent newspaper article snippets (Sec. 4).

Summary of our contributions:

1. continuously evolving embeddings:
   - simple and intuitive method for computing diachronic embeddings
   - can be applied to small datasets thanks to pre-trained transformer models
   - corpus does not need to be split into (arbitrarily) defined time intervals
   - frequent snapshots ensure smoothly changing, high-resolution embeddings
2. all the necessary code[1] to explore word usage change in novel datasets with a user-friendly web app

2 Continuously Evolving Embeddings

Let $\mathbf{x}_{t_i}^{\text{local}}$ be the contextualized embedding of a token t generated by some arbitrary method (e.g. a pre-trained BERT model) for the i^{th} occurrence of t in a corpus. Then a global embedding of t can be computed by averaging over the local embeddings of all N occurrences of t in the corpus (Horn, 2017; Bommasani et al., 2019; Martinc et al., 2019; Kutuzov and Giulianelli, 2020):

$$\mathbf{x}_t^{\text{global}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_{t_i}^{\text{local}}. \tag{1}$$

Equivalently, this can be formulated as a running average (Finch, 2009), allowing for memory-efficient continuous updates in one pass over the corpus:

$$\mathbf{x}_t^{\text{global}[:n]} = \frac{(n-1)\,\mathbf{x}_t^{\text{global}[:n-1]} + \mathbf{x}_{t_i}^{\text{local}}}{n}.$$

Using this running average formula, it is possible to compute continuously evolving embeddings by updating the global embedding as more and more sentences are processed (Akbik et al., 2019). However, usually the more recent occurrences of the word are of greater relevance when determining the current sense of the word. To account for this, the above formula can be adapted by introducing a weighting factor $0 < \alpha \leq 0.5$:

$$n' = \min\left\{n, \left\lceil \frac{1}{\alpha} \right\rceil \right\}$$

$$\mathbf{x}_t^{\alpha[:n]} = \frac{(n'-1)\,\mathbf{x}_t^{\alpha[:n-1]} + \mathbf{x}_{t_i}^{\text{local}}}{n'}. \tag{2}$$

This is equivalent to computing the weighted average of the two embeddings (for large n):

$$\mathbf{x}_t^{\alpha[:n]} = (1-\alpha)\,\mathbf{x}_t^{\alpha[:n-1]} + \alpha\,\mathbf{x}_{t_i}^{\text{local}}$$

and results in an exponential forgetting of the past occurrences in favor of the more recent instances (Finch, 2009).

While Martinc et al. (2019) generate diachronic embeddings by computing a global average of all contextualized embeddings occurring in texts from individual (predefined) time periods (Eq. 1), we instead propose to keep track of a weighted running average computed in one pass over the whole (chronologically ordered) corpus (Eq. 2). By taking ‘snapshots’ of the current state of these continuously evolving embeddings at user-defined time points, it is possible to obtain smoothly changing high-resolution diachronic word embeddings.[2] The weighting parameter α in the running average should be set according to the number of word occurrences one assumes it might take for the meaning to change and can be set individually for each word to reflect the differing overall frequencies and semantic shift paces (Hamilton et al., 2016).

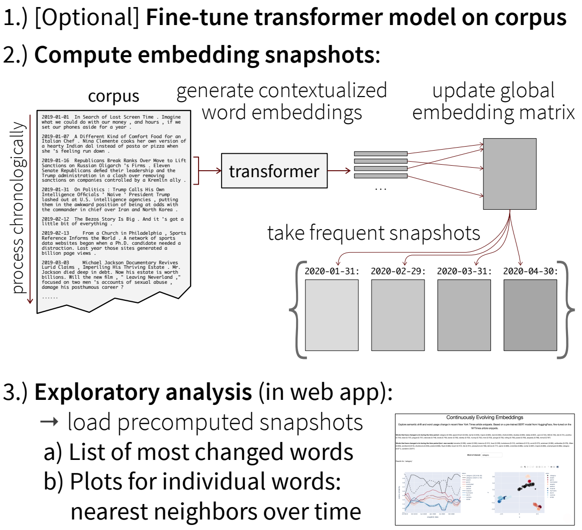
The computation of continuously evolving embeddings scales linearly with respect to the number of sentences in the dataset, since each sentence has to be embedded with the transformer model once to update the weighted running average with the respective contextualized embeddings. The required memory, on the other hand, scales linearly with the number of embedding snapshots that are taken during the computation, where a copy of the current state of the global embedding matrix needs to be stored for every snapshot.

[1] https://github.com/cod3licious/evolvemb

[2] Since an embedding simply stays the same when the word does not occur, these snapshots can be taken in arbitrarily short intervals.

3 The EvolvEmb App

Word usage changes in a corpus can be easily explored using the web application we created for this purpose. The app itself is based on the dash framework (Shammamah Hossain, 2019) and can be run locally by following the steps listed in Fig. 2 and demonstrated in the screencast[3], i.e., first computing continuously evolving embeddings and saving the respective snapshots (or, alternatively, diachronic embeddings obtained with a traditional approach such as a SGNS model trained on individual time slices (Kim et al., 2014)), and then starting the app (which loads the pre-computed embeddings) to obtain the list of most changed words in the corpus and a simple interface to generate the plots displaying the evolution of nearest neighbors over time for individual (user-selected) words.


1.) [Optional] **Fine-tune transformer model on corpus**
2.) **Compute embedding snapshots**:
corpus
generate contextualized word embeddings
update global embedding matrix
process chronologically
transformer
...
take frequent snapshots
2020-01-31:
2020-02-29:
2020-03-31:
2020-04-30:
3.) **Exploratory analysis** (in web app):
→ load precomputed snapshots
a) List of most changed words
b) Plots for individual words: nearest neighbors over time

```
2019-01-01  In Search of Lost Screen Time . Imagine
what we could do with our money , and hours , if we
set our phones aside for a year .

2019-01-07  A Different Kind of Comfort Food for an
Italian Chef . Nina Clemente cooks her own version of
a hearty Indian dal instead of pasta or pizza when
she 's feeling run down .

2019-01-16  Republicans Break Ranks Over Move to Lift
Sanctions on Russian Oligarch 's Firms . Eleven
Senate Republicans defied their leadership and the
Trump administration in a clash over removing
sanctions on companies controlled by a Kremlin ally .

2019-01-31  On Politics : Trump Calls His Own
Intelligence Officials ' Naive ' President Trump
lashed out at U.S. intelligence agencies , putting
them in the awkward position of being at odds with
the commander in chief over Iran and North Korea .

2019-02-12  The Bezos Story Is Big . And it 's got a
little bit of everything .

2019-02-13     From a Church in Philadelphia , Sports
Reference Informs the World . A network of sports
data websites began when a Ph.D. candidate needed a
distraction. Last year those sites generated a
billion page views .

2019-03-03     Michael Jackson Documentary Revives
Lurid Claims , Imperiling His Thriving Estate . Mr.
Jackson died deep in debt. Now his estate is worth
billions. Will the new film , " Leaving Neverland ,"
focused on two men 's accounts of sexual abuse ,
damage his posthumous career ?

......
```

Figure 2: Steps to explore word usage change in novel datasets: First the continuously evolving embedding snapshots are computed as described in Sec. 2, then the precomputed matrices can be used in the app to produce interactive plots similar to those shown in Fig. 3.

4 Exploring Word Usage Change

To demonstrate our approach, we downloaded 95,203 newspaper article snippets (consisting of a headline and 1-3 sentences) published by the New York Times between April 1st, 2019, and Dec. 31st, 2020, via their API.[4] Diachronic embeddings were computed for the 5,620 words that occurred at least 50 times in the corpus by processing the texts chronologically, computing continuously evolving embeddings with a transformer model, and taking a snapshot of the current state of the embeddings at the end of each month. α was set individually for each word based on how many times on average the word occurred in the articles of a single month. To compute the contextualized embeddings, we experimented with pre-trained BERT and RoBERTa models from the HuggingFace library (Wolf et al., 2020) that were either used as is or fine-tuned for three epochs on our corpus. As the results obtained with both models were similar, we focus on BERT in the following.

Words with different usages were identified based on the minimum cosine similarity between their embedding snapshots from different time points.[5] As this also yielded several words with multiple meanings that showed seasonal trends (Table 1, Fig. 3), we additionally identified words with a continuous semantic shift specifically by considering only the cosine similarity scores S_{ik} of all snapshots i to the last snapshot k and subtracted from the overall increase of the scores over time any intermediate decrease between subsequent scores:

$$(S_{kk} - S_{0k}) - \sum_{i=0}^{k-1} \max\{S_{ik} - S_{(i+1)k}, 0\}.$$

As expected, when computing continuously evolving embeddings on shuffled article snippets, i.e., a corpus that is no longer chronologically ordered (Dubossarsky et al., 2017), the resulting semantic shift scores are significantly lower (Table 2).

Even though the continuously evolving embeddings computed with pre-trained transformers are already sufficient to identify many words with usage changes, fine-tuning of the models is generally advised, especially to clearly identify semantic shifts when the new usage of a word was not present in the texts the transformer was originally trained on. To illustrate this, inspired by Rosenfeld and Erk (2018), we introduced a synthetic semantic shift into the data for an artificially created

[3] https://youtu.be/ltF67J-la7I

[4] https://developer.nytimes.com/apis

[5] We also explored the intersection of the k nearest neighbors (NN) of the word embeddings to identify words with a usage change (Gonen et al., 2020), but found these results to be less reliable, because the number of close NN can differ a lot between words with a specific or more general meaning. Nevertheless, there is a significant correlation between the minimum cosine similarity and the kNN interaction score.

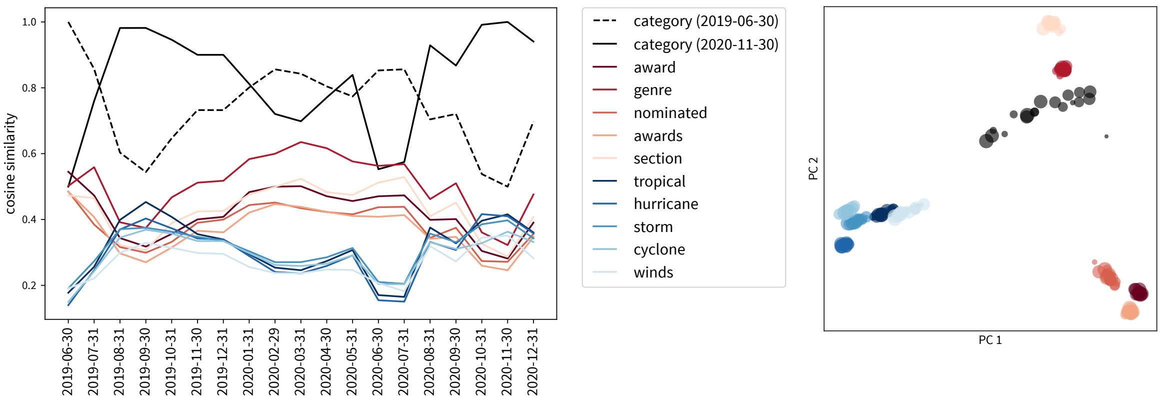


Figure 3: Plots as included in the app, here depicting the evolution of nearest neighbors over time for the word "category", computed with a pre-trained BERT model on our NYTimes article snippets dataset. For the target word, first the two time points with the smallest cosine similarity between the embeddings of the word itself were identified, then the five nearest neighbors of the word at both time points were selected (red and blue colors respectively; words that occurred in both sets are in red). *Left:* Cosine similarity between the target word at each time point and the nearest neighbors, as well as the two most different embedding snapshots of the target word itself (inspired by the plots in (Bamler and Mandt, 2017)). *Right:* 2D PCA visualization of all embedding snapshots of the target word as well as both sets of nearest neighbors (smaller dots represent embeddings at earlier time points).

Table 1: The 25 most changed tokens with their corresponding minimum cosine similarity score between the embedding snapshots *(multiple meanings)* and our semantic shift score, obtained by computing continuously evolving embeddings using a pre-trained BERT model on the NYTimes article snippets (ignoring new words that only occurred after the first snapshot date; words occurring in both lists are italicized).

multiple meanings: category (0.50), appointment, *barrier*, majors, bend, chiefs, doubles, tables, upon, *600*, del, *positive*, *kobe*, *plague*, nationals, lands, *dorian*, stanley, murray, mine, *plunge*, rolling, posed, jeopardy, revival (0.77)
semantic shift: coney (0.1869), *kobe* (0.1852), *dorian*, *600*, *barrier*, *plague*, stimulus, remotely, arbery, *positive*, sheet, thanksgiving, excerpt, tudor, *plunge*, halted, mask, infected, tracing, distancing, masks, educators, throwing, tip, retire (0.1083)

Table 2: Analogous to Table 1: the 25 tokens with the greatest semantic shift when applying the pre-trained BERT model to shuffled sentences.

semantic shift (shuffled): breakthrough (0.1210), trend (0.0621), coup, urgency, releasing, succeed, wind, limiting, holes, forecast, developments, attempted, richest, superstar, pastor, addressing, pack, upset, recommendation, programming, autism, arrival, denver, associated, flowers (0.0313)

word:[6] First, we removed all sentences containing the words "president" or "coronavirus" (the two most frequent nouns in our dataset) from the corpus and replaced each occurrence of the respective word with the new token "presidentcoronavirus". These augmented sentences were then reintroduced into the corpus at regular intervals based on a transition probability that follows a sigmoid curve, i.e., most of the sentences included at earlier dates were sampled from the contexts for the word "president", while at later dates this shifted towards sentences that originally contained "coronavirus". While the continuously evolving embeddings computed with a pre-trained BERT model can pick up on this artificially introduced semantic shift in general (Fig. 4 *top*: the black lines for the token 'presidentcoronavirus' run according to the sigmoid curve based on which the respective contexts were sampled), the nearest neighbors are not very instructive to identify the two senses. This is mainly due to the subword embeddings that the transformer uses to represent this new token, thereby introducing a strong preconception w.r.t. the word's meaning. However, after fine-tuning BERT on the synthetic dataset for three epochs, not only is the difference between the embedding snapshots of the target to-

[6] Since established word usage change evaluation datasets so far only cover broad discrete time bins (Schlechtweg et al., 2020), to evaluate gradual semantic shifts one has to resort to synthetic data (Shoemark et al., 2019).

ken itself stronger, but also the nearest neighbors now correspond more closely to the initial ('president') and later ('coronavirus') sense of the word (Fig. 4 *bottom*).

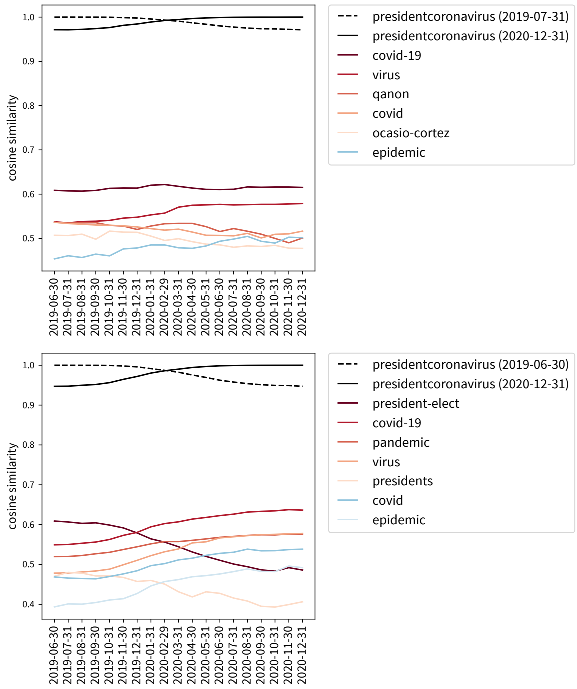


Figure 4: Analogously to Fig. 3 the nearest neighbors over time for the artificially constructed token "presidentcoronavirus" before (*top*; semantic shift score: 0.028) and after (score: 0.053) fine-tuning BERT on the synthetically modified NYT article snippets.

Finally, as a comparison we also show the plots obtained with diachronic embeddings learned using a skip-gram word2vec model trained with negative sampling (SGNS) on the original sentences. Similar to Kim et al. (2014), we trained a SGNS model[7] from the gensim library (Řehůřek and Sojka, 2010) for 50 epochs on the texts from each time period between two snapshots. As described in the original paper, the embeddings learned on later time slices were initialized with the embeddings from the previous interval. Additionally, since the amount of text contained in each time slice is much smaller than generally recommended when training a word2vec model, the model was first trained on the full corpus for 100 epochs to initialize the embeddings before training on the first time period. While the evolution of nearest neighbors over time (Fig. 5) still contains faint patterns (e.g., the sense "hurricane" is stronger during the late summer and fall months), the plots are much noisier than those created with the transformer-based continuously evolving embeddings (Fig. 3).

5 Related Work

When it comes to learning word embeddings in general, it is helpful to distinguish between older methods, such as word2vec (Mikolov et al., 2013a,b) or GloVe (Pennington et al., 2014), that learn static word embeddings, i.e., a single "global" embedding for each word in the vocabulary, and modern transformer models, such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and Flair (Akbik et al., 2018), that generate contextualized embeddings based on the local context of a word in the current sentence. While the static word embedding models are usually trained on a target corpus containing several millions of words to obtain expressive domain-specific embeddings (Tshitoyan et al., 2019), pre-trained transformers are well suited for transfer learning and can therefore also more easily be applied to smaller datasets.

Some of the more advanced methods for creating diachronic embeddings use special-purpose dynamic language models, which explicitly take the temporal structure of the data into account when learning the word embeddings (Bamler and Mandt, 2017; Rosenfeld and Erk, 2018; Yao et al., 2018; Rudolph and Blei, 2018; Brandl and Lassner, 2019; Jawahar and Seddah, 2019; Hofmann et al., 2020; Tsakalidis and Liakata, 2020). A different line of work instead relies on conventional static word embedding models, such as word2vec, and uses them directly to learn embeddings for the individual time periods. The main challenge here consists of aligning the word embeddings learned for different time intervals, which can, for example, be achieved by using the embeddings from one time slice to initialize the next (Kim et al., 2014), by explicitly matching up the matrices learned on different time periods (Kulkarni et al., 2015; Hamilton et al., 2016; Zhang et al., 2016; Yin et al., 2018), or utilizing other techniques such as temporal referencing (Dubossarsky et al., 2019). An even simpler approach to produce diachronic word embedding in a single embedding space uses the same (fixed) model to compute contextualized embeddings on all texts and then averages the respective embeddings from each individual time period to get the diachronic embeddings (Basile et al.,

[7] embedding dim. 50; context window 5; neg. sampling 13

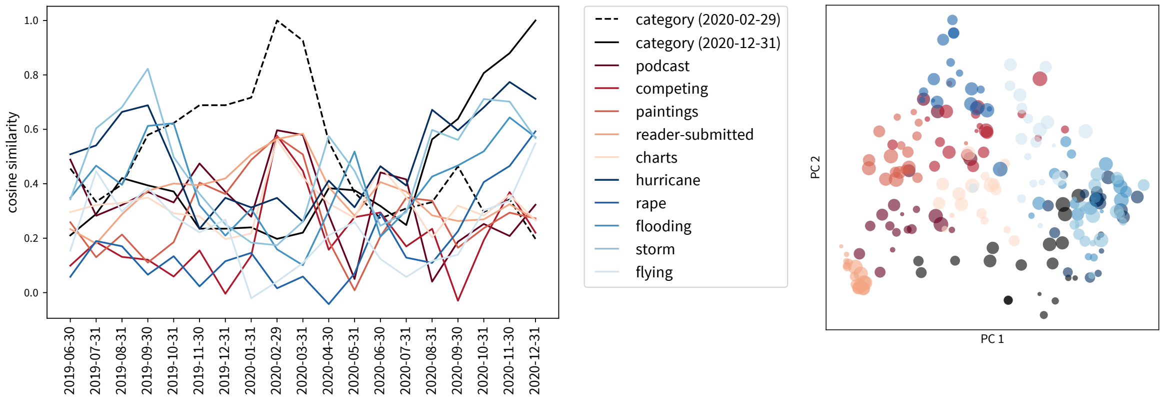


Figure 5: Analogously to Fig. 3 the results with diachronic embeddings obtained by training a SGNS word2vec model on the article snippets from the respective time intervals (Kim et al., 2014).

2016).[8] When using high-quality contextualized embeddings from a transformer model, it is furthermore possible to compute the diachronic embeddings for shorter time slices of single years (Martinc et al., 2019, 2020; Hu et al., 2019; Giulianelli et al., 2020; Beck, 2020). However, one major problem remains, namely that the time slices across which the diachronic embeddings are computed have to be discretized and defined in advance. This problem could previously only be addressed by a more complex dynamic model (Rosenfeld and Erk, 2018).

While several of the above mentioned papers have published code alongside their manuscripts, this was mainly done with the intention that others could reproduce their results, not apply the methods to novel datasets. To the best of our knowledge, only Hamilton et al. (2016) has released a more comprehensive library to explore word usage change in other corpora, however, their approach relies on static word embeddings and should therefore mainly be applied to larger corpora. Most other available software for analyzing corpora only considers word frequencies over time, but does not track the semantic shifts of these words.

6 Conclusion

This paper introduced continuously evolving embeddings as a conceptually simple and intuitive method for computing smoothly changing high-resolution diachronic embeddings from weighted running averages of contextualized embeddings. By taking advantage of pre-trained transformer models and processing the texts in a corpus sequentially rather than dividing them into (more or less arbitrary) time slices, our approach makes it possible to obtain diachronic embeddings from comparatively small corpora and at very short intervals compared to the previously standard time periods of at least one year. This should make our method particularly well suited to study fast-paced environments such as social media, where a new meme can go viral in a matter of hours, only to be superseded by the next a few days later.

Aside from the parameters involved in the underlying transformer model and its possible fine-tuning, our method only has a single hyperparameter, α, whose setting mostly just influences how frequently the embedding snapshots need to be taken to not miss any semantic shifts in between the snapshot intervals. On our NYTimes corpus we obtained reasonable results already with pre-trained transformer models, however, fine-tuning is nevertheless advised and especially helpful to characterize new word usages that the transformer did not encounter in its original training data.

We hope that the provided code will help others identify interesting patterns of word usage change in their own corpora.

Acknowledgments

This paper would not be what it is today without the many fruitful discussions with Alan Akbik and his support. I would like to thank him, Klaus-Robert Müller, and the anonymous reviewers for their helpful comments on this manuscript. FH acknowledges funding from the BZML.

[8] Since the contextualized embeddings are all in the same embedding space already (defined by the single fixed model), averaging the embeddings from each time slice creates time period specific global word embeddings that are themselves also comparable.

References

Alan Akbik, Tanja Bergmann, and Roland Vollgraf. 2019. Pooled contextualized embeddings for named entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 724–728.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 380–389.

Pierpaolo Basile, Annalina Caputo, Roberta Luisi, and Giovanni Semeraro. 2016. Diachronic analysis of the italian language exploiting google ngram. *CLiC it*, pages 56–60.

Christin Beck. 2020. DiaSense at SemEval-2020 Task 1: Modeling sense change via pre-trained BERT embeddings. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 50–58.

Andreas Blank. 1999. Why do new meanings occur? a cognitive typology of the motivations for lexical semantic change. *Historical semantics and cognition*, 13(6).

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2019. Bert wears gloves: Distilling static embeddings from pretrained contextual representations.

Stephanie Brandl and David Lassner. 2019. Times are changing: Investigating the pace of language change in diachronic word embeddings. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 146–150.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Haim Dubossarsky, Simon Hengchen, Nina Tahmasebi, and Dominik Schlechtweg. 2019. Time-out: Temporal referencing for robust modeling of lexical semantic change. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 457–470.

Haim Dubossarsky, Daphna Weinshall, and Eitan Grossman. 2017. 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.

Tony Finch. 2009. Incremental calculation of weighted mean and variance. *University of Cambridge*, 4(11-5):41–42.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*.

Valentin Hofmann, Janet B Pierrehumbert, and Hinrich Schütze. 2020. Dynamic contextualized word embeddings. *arXiv preprint arXiv:2010.12684*.

Franziska Horn. 2017. Context encoders as a simple but powerful extension of word2vec. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 10–14. Association for Computational Linguistics.

Shammamah Hossain. 2019. Visualization of Bioinformatics Data with Dash Bio. In *Proceedings of the 18th Python in Science Conference*, pages 126 – 133.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908.

Ganesh Jawahar and Djamé Seddah. 2019. Contextualized diachronic word representations. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 35–47.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal analysis of language through neural language models. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 61–65.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, pages 625–635.

Andrey Kutuzov and Mario Giulianelli. 2020. Uio-uva at semeval-2020 task 1: Contextualised embeddings for lexical semantic change detection. *arXiv preprint arXiv:2005.00050*.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020. Capturing evolution in word usage: Just add more clusters? In *Companion Proceedings of the Web Conference 2020*, pages 343–349.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2019. Leveraging contextual embeddings for detecting diachronic semantic shift. *arXiv preprint arXiv:1912.01072*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP*, volume 14, pages 1532–1543.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Alex Rosenfeld and Katrin Erk. 2018. Deep neural models of semantic shift. 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 474–484.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference*, pages 1003–1011.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23.

Philippa Shoemark, Farhana Ferdousi Liza, Dong Nguyen, Scott Hale, and Barbara McGillivray. 2019. Room to glo: A systematic comparison of semantic change detection approaches with word embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 66–76.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to lexical semantic change. *arXiv preprint arXiv:1811.06278*.

Adam Tsakalidis and Maria Liakata. 2020. Sequential modelling of the evolution of word representations for semantic change detection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8485–8497.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

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*, pages 673–681.

Zi Yin, Vin Sachidananda, and Balaji Prabhakar. 2018. The global anchor method for quantifying linguistic shifts and domain adaptation. *Advances in neural information processing systems*, 31:9412–9423.

Yating Zhang, Adam Jatowt, Sourav S Bhowmick, and Katsumi Tanaka. 2016. The past is not a foreign country: Detecting semantically similar terms across time. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2793–2807.