DCC-Uchile at SemEval-2020 Task 1: Temporal Referencing Word Embeddings

Frank D. Zamora-Reina
Department of Computer Science
University of Chile & IMFD
fzamora@dcc.uchile.cl

Felipe Bravo-Marquez
Department of Computer Science
University of Chile & IMFD
fbravo@dcc.uchile.cl

Abstract

We present a system for the task of unsupervised lexical change detection. Given a target word and two corpora spanning different periods of time, automatically detects whether the word has lost or gained senses from one corpus to another. Our system employs the temporal referencing method to obtain compatible representations of target words in different periods of time. This is done by concatenating corpora of different periods and performing a temporal referencing of target words i.e., treating occurrences of target words in different periods as two independent tokens. Afterwards, we train word embeddings on the joint corpus and compare the referenced vectors of each target word using cosine similarity. Our submission was ranked 6th among 33 teams for subtask 1, obtaining an average accuracy of 0.637, only 0.050 points behind the first ranked system.

1 Introduction

In the past decade we have seen a rise in academic work on the automatic detection of lexical semantic change (Tahmasebi et al., 2018; Kutuzov et al., 2018). However, many of these studies are difficult to compare because of differences in evaluation procedure, languages, corpora, and time periods.

SemEval 2020 Task 1 aims to introduce a simple evaluation framework for unsupervised lexical semantic change detection (Schlechtweg et al., 2020). The task focuses on four languages: German, English, Swedish, and Latin. For each language, two corpora (C_1 and C_2) are provided, each of those spanning different periods of time (t_1 and t_2).

The task is divided into two subtasks described below.

1. Subtask 1 - Binary classification: for a set of targets words, decide which words lost or gained senses between t_1 and t_2; as annotated by human judges.

2. Subtask 2 - Ranking: rank the same set of target words according to their degree of lexical semantic change between t_1 and t_2 (the stronger the change the higher the position in the ranking).

Systems are evaluated against ground truth data annotated by human native speakers (except for Latin, which was annotated by scholars of Latin) (Schlechtweg et al., 2018).

In this paper, we present our system for detecting semantic change of words in different periods of time. Our system was solely designed for subtask 1.

The rationale of our approach is to obtain compatible representations of target words in different periods of time. This is done by concatenating corpora of different periods and performing a temporal disambiguation of target words. Afterwards, we train word embeddings on the joint corpus and compare the disambiguated vectors of each target word using cosine similarity. This process is known as temporal referencing and was proposed in (Dubossarsky et al., 2019).

Our submission was ranked 6th among 33 teams for subtask 1, obtaining an average accuracy of 0.637, only 0.050 points behind the first ranked system. Temporal referencing is a very simple technique that despite its simplicity yielded competitive results and outperformed baselines across all languages.

2 Related Work

Linguists are well aware that language, even when viewed as a synchronous entity, is full of variation at all linguistic levels. In spoken language, this kind of variation is the norm.

The detection of lexical change is complex due to the fact that the variation observed in lexical form between different corpora is not necessarily due to diachronic causes, as it may also be due to other factors such as the presence of regional dialects.

One of the most challenging aspects of comparing word embeddings trained on different periods of time is alignment, which is the process of making embeddings models trained on different corpora, compatible and comparable. Specifically, most cost functions for training word representations are invariant to rotations. Consequently, the learned embeddings of independent models i.e., models trained from corpora spanning different periods of time, may not be placed in the same latent space. This problem is known as the alignment problem. In general, most of the previous approaches to studying lexical change follow a similar two-step methodology: first compute static word embeddings in each separate period of time, and then find a way to align the word embedings across different time periods (Yao et al., 2018).

Dubossarsky et al. (2019) proposed the *Temporal Referencing* method for lexical semantic change on which this work is based. A single embedding model is trained on a corpus that combines various diachronic corpora in which the target words (i.e., words on which semantic change will be assessed) are disambiguated according to the corpus in which they occur. Two models are compared: PPMI (Levy et al., 2015) and Skip-gram negative sampling (SGNS) (Mikolov et al., 2013b), in two settings: 1) using alignment and 2) using temporal referencing. Results showed that the SGNS model trained with temporal referencing exhibits significantly less noise than its aligned counterpart, and for the PPMI model, temporal referencing significantly reduced the noise level, but to a lesser extent than for SGNS.

The work of Gulordava and Baroni (2011) is based on context vectors to identify semantic change over time. A co-occurrence matrix with local mutual information (LMI) scores is built (Evert, 2008). Then, the distributional similarity of each word pair is measured as the cosine product of the corresponding context vectors. A partitions of Google Book Ngram corpus spanning from 1960s and 1990s is used for evaluation. The main finding of the study is that the proposed method can successfully detect the semantic change of words, as well as cases of major diachronic context change.

Kim et al. (2014) developed a method employing prediction-based word embedding models to trace diachronic semantic shifts. They used incremental updates and Skip-gram with negative sampling (SGNS) (Mikolov et al., 2013a) as the embedding model. The authors identified words that have changed and the periods during which they changed.

In (Bamler and Mandt, 2017), the authors propose the dynamic skip-gram model, which is a Bayesian probabilistic model that combines the skip-gram architecture with a latent continuous time series. The method is jointly trained over multiple time period and does not require an alignment step. The experimental results showed that both dynamic skip-gram filtering (which conditions only on past observations) and dynamic skip-gram smoothing (which uses all data) lead to smoothly changing embedding vectors that are better at predicting word-context statistics at held-out time steps.

In (Yao et al., 2018), a dynamic statistical model was developed to learn time-aware word vector representations. This model can simultaneously learn time-aware embeddings avoiding the need for alignment. This model is trained on a corpus of New York Times articles. Additionally, multiple evaluation strategies of temporal word embeddings are conducted. The qualitative and quantitative tests indicate that the proposed method not only reliably captures the evolution of words over time, but also consistently outperforms state-of-the-art temporal embedding approaches on both semantic accuracy and alignment quality.

Another way of creating dynamic embedding was developed in (Rudolph and Blei, 2018) based on embeddings of exponential families and latent variables with a random walk Gaussian drift. The key idea is to restrict the embedding vectors to a single a time slice (Tahmasebi et al., 2018) while sharing the context vectors across all time points. The authors experimented with three diachronic corpora: 1) ArXiv's machine learning papers (2007–2015), 2) ACM's computer science abstracts (1951–2014), and 3) U.S. Senate speeches (1858–2009). The results from this study suggest that dynamic embeddings

provide a better fit than their static counterparts for working with diachronic corpora and that dynamic embeddings can also help identify ways in which language changes.

3 System Description

Our system implements the *Temporal Referencing* method (Dubossarsky et al., 2019), which is based on the idea of obtaining compatible representations of target words in different periods of time. This is done by concatenating the two corpora C_1 and C_2 coming from different periods and performing a temporal referencing of target words. It is important to remark that our system was designed only to participate in subtask 1.

Our hypothesis is that if a word exhibits a semantic change over time then it representation should change too. However, if we independently train word representation for each period of time, our vectors would not be comparable (because of the stochasticity of neural networks). One approach would be to perform an alignment on the different embedding models. In contrast, we concatenate for each language the two temporal corpora (C_1 and C_2) and referenced target words by appending special temporal symbols to their tokens. Next, we describe our own implementation of the temporal referencing method. The process is illustrated in Figure 1.

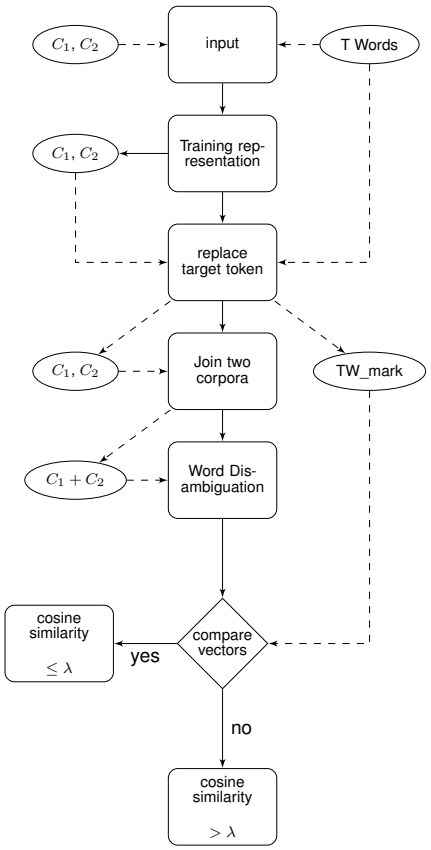

Figure 1: Temporal referencing workflow.

The first step is to merge both temporal corpora for each language. Then, for each target word, we add a special symbol to all its appearances in both corpora, the symbol denoting the corresponding corpus or period. For example, the target word *"attack"* becomes *"*$attack\_t1$*"* in C_1 and *"attack_t2"* in C_2.

Afterwards, we proceed to merge both corpora, and in this way old and new occurrences of a target word occur in contexts of words coming from a single vocabulary of non-target terms.

The next step consists of training a distributed representation of the words from the joined corpus. We use the Continuous Bag of Words (CBOW) model with negative sampling to obtain our word vectors (Mikolov et al., 2013a). Consequently, our system does not need to align the word embeddings across different time spans.

Finally, in order to classify whether a target word exhibits semantic change, we calculate the cosine similarity between the two vectors of each target word. We use a threshold parameter λ to convert the similarities into binary classes 1 (semantic change) and 0 (no semantic change) [1]. The value of λ is determined experimentally.

4 Experiments

4.1 Data

We carried out our experiments on the SemEval 2020 Task 1 dataset. This dataset is made of pairs of corpora spanning two different time periods, with one pair for each language: English, German, Latin and Swedish. The time periods covered in each corpus are shown in Table 1. The number of tokens and target words for each corpus are shown in Table 2.

Languages	t1	t2
English	1810-1860	1960-2010
German	1800-1899	1946-1990
Latin	-200-0	0-2000
Swedish	1790-1830	1895-1903

Table 1: Time period covered by each temporal corpus and language.

Languages	#tokens-t1	#tokens-t2	Target words
English	6 million	6 million	37 lemmas
German	70 million	72 million	48 lemmas
Latin	1.2 million	9.4 million	40 lemmas
Swedish	71 million	111million	31 lemmas

Table 2: Number of tokens for each temporal corpus and language together with the number of target words for each language.

4.2 Parameters Settings

We used the *Gensim*[2] package for training our word2vec embeddings. The number of dimensions was set to 100, the window size was set to 5, and the learning model we used was Continous Bag of Words (CBOW) with negative sampling. The number of negative samples was set to 5. The classification threshold of the cosine similarity was set to $\lambda = 0,61$. This value was intuitively adjusted due to the lack of validation data.

4.3 Baselines

The shared task provides two baseline models described below.

1. Normalized frequency difference (FD): This method counts the frequency of each target word in each of the two corpora, normalizes these counts by number of words of the corresponding corpus, and then calculates the absolute difference of these values as a measure of change.

2. Count vectors with column intersection and cosine distance (CNT+CI+CD): this method learns word representations for the two corpora, then aligns the vectors by intersecting their columns and measures change using the cosine similarity between the two vectors of a target word.

[1] Code is available at `https://github.com/fdzr/experiments_semeval`

[2] `https://radimrehurek.com/gensim/`

4.4 Results

Our system was evaluated by contrasting our binary predictions against the manually annotated gold labels using accuracy (ACC) as evaluation metric.

The classification performance of our model and the two baselines are shown in Table 3.

System	English	German	Latin	Swedish	Avg
baseline 1	0.432	0.417	0.650	0.258	0.439
baseline 2	0.595	0.688	0.525	0.645	0.613
our system	0.649	0.667	0.525	0.710	0.637

Table 3: Classification accuracy of baselines and our system. The best results per column are shown in gray.

The results show that our system outperformed both baselines across all languages. Moreover, we obtained the 6th place for subtask 1 out of 33 participants.

In relation to subtask 2, we made a random submission to meet the competition requirements as shown in Table 4. It is important to note that these results do not indicate the performance of the temporal referencing technique in the ranking subtask. We plan to investigate how to adapt our system to this subtask in future work.

System	English	German	Latin	Swedish	Avg
our system	-0.217	0.014	0.020	-0.150	-0.083

Table 4: Spearman correlation of our random submission to subtask 2.

4.5 Post-Evaluation Results

In this subsection we show experimental results conducted after the release of the gold data. The purpose of these experiments is to gain more insight into the effect of hyperparameters that we could not calibrate during the training phase due to the unsupervised nature of this task.

In particular, we experimented with two word embedding architectures CBOW and SGNS, and different values for the window and the embedding size. The number of negative samples was kept at 5 in all experiments. The results are depicted in Table 5.

Parameters			Languages				
architecture	window size	embedding size	English	German	Latin	Swedish	Avg
CBOW	2	100	0.568	0.688	0.375	0.742	0.593
	5		0.595	0.729	0.500	0.744	0.649
	10		0.676	0.750	0.475	0.710	0.653
	2	200	0.595	0.708	0.400	0.710	0.603
	5		0.649	0.771	0.500	0.774	0.673
	10		0.649	0.792	0.525	0.710	0.669
SGNS	2	100	0.568	0.646	0.350	0.710	0.568
	5		0.568	0.646	0.350	0.806	0.592
	10		0.622	0.688	0.400	0.774	0.621
	2	200	0.595	0.688	0.350	0.710	0.585
	5		0.595	0.667	0.400	0.774	0.609
	10		0.649	0.750	0.450	0.645	0.623

Table 5: Classification accuracy using two word2vec architectures and various parameter settings. The classification threshold λ was set to 0.5. The best results per column are shown in gray.

Our results indicate that the use of a short context window is not appropriate for this particular task. We also observed a tendency for CBOW to achieve better results than SGNS. Moreover, the results for Latin are notoriously worse than for any other language.

In order to better understand the effect of the decision threshold λ on classification performance, we conducted an additional experiment using different values for this parameter. These experiments are done

using the best configuration found in the previous experiment, which corresponds to a CBOW architecture with window size 5 and embedding size 200. The results are showed in Table 6.

λ	English	German	Latin	Swedish	Avg
0.4	0.595	0.750	0.375	0.774	0.623
0.5	0.649	0.771	0.500	0.774	0.673
0.6	0.676	0.646	0.525	0.742	0.647
0.7	0.703	0.458	0.525	0.548	0.559
0.8	0.514	0.354	0.700	0.258	0.456

Table 6: Classification accuracy using different values of λ for the best configuration of the previous experiment. The best results per column are shown in gray.

We observe that the value of λ produces significant variability in accuracy across the four languages. Surprisingly, lower values of λ are good for German and Swedish but bad for Latin. This suggests that this parameter needs to be adjusted independently for each particular language. Furthermore, this also suggests that evaluation metrics that do not depend on the decision threshold, such as the area under the ROC curve (AUROC), may be a more appropriate choice for this task.

5 Conclusions

We presented a system for the automatic detection of semantic change in words occurring in two different periods of time. Our system implements the temporal referencing method (Dubossarsky et al., 2019) in which compatible representations of target words are obtained and compared by concatenating corpora of different periods and performing a temporal disambiguation of target words. Our submission, in addition to outperforming both baselines, ranked 6th place in the binary classification subtask of SemEval 2020 Task 1. These results provide further evidence that temporal referencing is a strong and simple approach for lexical semantic change detection.

For future work we will explore using representations of words obtained from deep contextualized models such as ELMO (Peters et al., 2018) and BERT (Devlin et al., 2019). We will also use supervised approaches for learning a similarity metric for our target words trained on large synthetically generated data.

6 Acknowledgments

This work was funded by Millennium Institute for Foundational Research on Data.

References

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

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, June. 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. pages 457–470, July.

Stefan Evert. 2008. Corpora and collocations. *Corpus linguistics. An international handbook*, 2:1212–1248.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the google books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71.

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, Baltimore, MD, USA, June. Association for Computational Linguistics.

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, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

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.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics*. ACL.

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, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana, June. Association for Computational Linguistics.

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 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of Computational Approaches to Lexical Semantic Change. *arXiv preprint arXiv:1811.06278*.

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.