HSE at TempoWiC: Detecting Meaning Shift in Social Media with Diachronic Language Models

Elizaveta Tukhtina, Svetlana Vydrina, Kseniia Kashleva
HSE University
{eatukhtina, svvydrina, kkashleva}@edu.hse.ru

Abstract

This paper describes our methods for temporal meaning shift detection, implemented during the TempoWiC shared task. We present two systems: with and without time span data usage. Our approach is based on masked language models continuously pre-trained with Twitter data. Both systems outperformed all the competition's baselines except TimeLMs-SIM. Our best submission achieved the macro-F1 score of 70.09% and took the 7th place. This result was achieved by using diachronic language models from the TimeLMs project.

1 Introduction

It is a commonplace that words change their meanings and connotations through time. Despite numerous studies about that, there are still difficulties in semantic change detection. Static embeddings are not suitable for working with semantic change, since they cannot reflect the fact that a word can have completely unrelated meanings. In this work, we are focusing on contextualized embeddings, which produce different vector representations depending on the context.

There were a number of competitions dedicated to semantic change detection, for example, TempoWiC (Loureiro et al., 2022b) and LSCDiscovery (Kashleva et al., 2022). These competitions were aimed at determining the difference in the meanings of words depending on the time period in which they are used. Datasets at LSCDiscovery consisted of texts from different centuries (Zamora-Reina et al., 2022). For this shared task, TempoWiC, the data with a time interval only of one year is used. It significantly changes the approach to the competition.

Temporal word in context (TempoWiC) benchmark aims to decide if there is a change between the meaning of two words in a given pair of tweets. TempoWiC is designed as a binary classification problem where the target word is featured in two tweets from different time periods, and the goal is to detect whether there is a meaning shift or not.

When creating a dataset for the competition, the authors decided to use data from social media (Twitter), while when developing the previous dataset WiC (Pilehvar and Camacho-Collados, 2018), word usage was taken from more formal sources such as Wiktionary, WordNet and VerbNet. The language used in social networks is much more informal and dynamic, so such a dataset is able to reflect even minor changes in word usage.

The TempoWiC dataset consists of paired tweets and is divided into train/validation/test samples of size 1,428/396/1,473 instances, respectively. For each sample, the set of target words is different. As additional data, the publication date is indicated for each tweet. There are no missing values in the dataset. Participants of the competition were asked to detect the change in the meaning of the target word both with and without using time-span information. To estimate the system's performance, the Macro-F1 score was used.

2 Systems Overview

In this section, we describe two systems that were implemented by our team during the shared task. Both systems are based on the pre-trained language models. For our first system, we did not use time-span information and extracted embeddings from the Twitter-roBERTa-base model (Barbieri et al., 2020). In the second approach, we tried to improve our system's performance by using diachronic language models from the TimeLMs project (Loureiro et al., 2022a).

2.1 General approach

First, we apply the continual learning strategy and train a masked language model with the TempoWiC data, using the script provided by the HuggingFace

Proceedings of the The First Workshop on Ever Evolving NLP (EvoNLP), pages 35 - 38

team[1]. Depending on the experiment, we use the entire dataset or a sub-set. In each experiment, we split a given corpus into train and test sets with a 95:5 ratio and train the model for 4 epochs with a learning rate of 2e-5. Then we extract summed representations for target words from all 12 layers of a corresponding language model. We decided to focus on word-level representations since they showed a better performance than sentence-level embeddings (see Appendix A for comparison results). For a word representation we adopt only the embedding of the first subword. Finally, we calculate cosine similarities between representations of two target words in each pair of tweets and use the obtained values to train a logistic regression model. Though the cosine-based approach is rather straightforward, its performance may strongly vary on the choice of the language model.

2.2 System 1. Twitter-roBERTa-base with additional pre-training

Our first system is based on the Twitter-roBERTa-base language model. It is a RoBERTa model (Liu et al., 2019) that was trained on 58M tweets. We chose this model because the competition's task was focused on Twitter data. For additional pre-training, we took all of the available tweets from the TempoWiC dataset, including train, validation, and test sets. For comparison, we also tried the BERT-base model (Devlin et al., 2018) as a baseline for our first system.

2.3 System 2. Diachronic models from the TimeLMs project

The distinguishing feature of the TempoWiC dataset is that each tweet has a specified time period: a year and a month when the tweet was posted. That means we can take into account not only the context of the tweet but also use time as an additional feature to improve our meaning shift detection system.

Our second solution is based on diachronic language models from the TimeLMs project. TimeLMs is a set of diachronic language models, based on RoBERTa, continuously trained on Twitter data over regular time intervals. The initial model was trained on tweets that were posted from 2018 until the end of 2019. Since the beginning of 2020, the base model has been continuously pre-training on diachronic Twitter data every three months. The project is active and at the time of this writing, models from 2019 to June 2022 are available. Since the TempoWiC dataset contains tweets from 2019, 2020, and 2021, models from the TimeLMs project can be applied to improve our first 'nondiachronic' approach.

Time span	Model
01.2019-12.2019	twitter-roberta-base-2019-90m
01.2020-03.2020	twitter-roberta-base-mar2020
04.2020-06.2020	twitter-roberta-base-jun2020
07.2020-09.2020	twitter-roberta-base-sep2020
10.2020-12.2020	twitter-roberta-base-dec2020
01.2021-12.2021	twitter-roberta-base-2021-124m

Table 1: TimeLMs models used in System 2 in accordance with time spans for the tweets from the TempoWiC dataset.

For our baseline diachronic approach, we used three TimeLMs models trained on Twitter data for a specific year: 2019, 2020 and 2021. The choice of the diachronic model for extracting the representation of a target word depends on the tweet's publication year. We also split the TempoWiC dataset into three corpora by year and additionally pre-trained each of the TimeLMs models with the corresponding corpus.

As an improvement strategy, we also decided to engage the TimeLMs models from a year's quarters. In this case, the choice of the model depends on the year's quarter in which the tweet was posted. Due to the lack of quarterly models for 2019 and because the number of tweets for 2021 was insufficient for pre-training quarterly models, this improvement was applied only to the tweets from 2020. Table 1 lists all of the TimeLMs models that were used for our second system.

3 Results

Table 3 shows the results for our two final systems. Submissions were evaluated using the Macro-F1 metric. Our best submission took 7th place. That is better than all baselines except TimeLMs-SIM (Logistic Regression based on Similarity of Contextual Embeddings from TimeLMs-2019-90M). This result is interesting because our best model (TimeLMs-with-quarter) seems more complex than the TimeLMs-SIM baseline, since we used a different model depending on the time span. According to the description of the baselines that became available after the evaluation phase, the task organiz-

[1]The script is available at https://github.com/huggingface/transformers/blob/main/examples/pytorch/language-modeling/run_mlm.py

Model	Validation (original LM)	Validation (LM with extra pre-training)	Test (original LM)	Test (LM with extra pre-training)
BERT-base-uncased	57.39	59.47	67.98	67.11
Twitter-RoBERTa-base	58.26	**67.61**	68.29	**68.76**
TimeLMs-by-year	57.50	66.63	63.81	68.69
TimeLMs-with-quarters	57.97	**68.70**	64.22	**70.09**

Table 2: Macro F1-scores for all of our models, including post-evaluation results for the Test set. Best results for Validation and Test sets are highlighted in bold.

ers used SP-WSD layer pooling weights (Loureiro et al., 2022d). Whereas in our system, we extracted summed embeddings from all 12 layers without pooling strategy.

		Submission 1	Submission 2
Rank	User/Baseline	Macro-F1	Macro-F1
Our results			
7	lisatukhtina	70.09	68.76
TOP-3 results from other teams			
1	dma	77.05	77.05
2	macd	76.60	74.74
3	zackchen	73.64	74.87
Baselines			
—	TimeLMs - SIM	70.33	
—	RoBERTa-L - SIM	67.09	
—	RoBERTa-L - FT	59.10	
—	TimeLMs - FT	57.70	
—	Random	50.00	
—	All True	26.79	

Table 3: Submission leaderboard

Table 2 shows detailed results for all the models we implemented during the shared task. For the validation set, there is a noticeable difference between the models with and without continued pre-training. We expected a similar trend for the test sample. To test this assumption, we obtained results for all our models in the post-evaluation phase. The results showed that for the BERT and Twitter-RoBERTa-base models, additional pre-training did not improve the quality on the test set. As for TimeLMs models, the results are correlated with the validation set. It is also interesting that such a general model as BERT-base performed as well as more complex solutions, even without pre-training for Twitter domain.

For the validation set, we also obtained Macro-F1 scores for each target word (see Table 4). The most challenging words for both models were *recount* and *primo*.

Word	Macro-F1	
	Twitter-RoBERTa-base	TimeLMs
impostor	65.97	70.62
lotte	67.07	64.10
recount	52.88	61.86
primo	57.65	60.17

Table 4: Macro-F1 scores for each word from the validation set.

4 Discussion

The main question that is still open is that: can we really detect a meaning shift on such a short time period as about a year? At TempoWiC it was postulated that in social media we can observe faster semantic shifts (Loureiro et al., 2022c). From a linguistic point of view, a change occurred, when there is evidence of transmission of innovations to others, i.e., of conventionalization (Traugott, 2017). It seems that one year is too short time for any language innovation to become widespread, even via social media. Moreover, it may take more time to be sure that this is a real change and not a nonce word. Let us consider words that were taken for a validation set. There were 4 of them (*lotte*, *primo*, *recount*, and *impostor*). According to the corpus provided by the organizers, the word *recount* was mostly used in the context of elections, *lotte* in the context of a concert and as a hotel name. It means that at least 50% of validation words demonstrate that trending words in Twitter in general most likely describe ongoing or recent events.

It was said that at TempoWiC the task was to decide if the meaning of the first target word in context is the same as the second one or not (Loureiro et al., 2022c). There are also examples of annotated sentences with target words in the article (see Table 5). These examples demonstrate polysemy, not semantic change. So it can be assumed that the TempoWiC dataset is much more suitable for word sense disambiguation task than for semantic change detection. It is difficult to differentiate

Tweet 1	Tweet 2	Label
2019-08 "In case you were wondering facial devotion still worked with a face *mask* on"	2020-08 "With these *mask* at work customers are forever confusing me and Reyna lmao"	1

Table 5: An example from the TempoWiC training set for a target word 'mask'. Label 1 indicates that the word has different meanings in the two tweets.

between polysemy and semantic change on such restricted data. That makes this shared task even more complicated.

5 Conclusion

We presented two systems for temporal meaning shift detection in Twitter, both with and without time span data usage. The best result was obtained with diachronic language models continuously trained for the Twitter domain. For our future research, we will consider weight-pooling methods as an attempt to improve our system's performance.

Acknowledgements

We deeply thank Lidia Pivovarova for her invaluable advice and help.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Kseniia Kashleva, Alexander Shein, Elizaveta Tukhtina, and Svetlana Vydrina. 2022. HSE at LSCDiscovery in Spanish: Clustering and profiling for lexical semantic change discovery. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 193–197, Dublin, Ireland. Association for Computational Linguistics.

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.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022a. Timelms: Diachronic language models from twitter.

Daniel Loureiro, Aminette D'Souza, Areej Nasser Muhajab, Isabella A. White, Gabriel Wong, Luis Espinosa Anke, Leonardo Neves, Francesco Barbieri, and Jose Camacho-Collados. 2022b. TempoWiC: An evaluation benchmark for detecting meaning shift in social media.

Daniel Loureiro, Aminette D'Souza, Areej Nasser Muhajab, Isabella A. White, Gabriel Wong, Luis Espinosa Anke, Leonardo Neves, Francesco Barbieri, and José Camacho-Collados. 2022c. Tempowic: An evaluation benchmark for detecting meaning shift in social media. *ArXiv*, abs/2209.07216.

Daniel Loureiro, Alípio Mário Jorge, and Jose Camacho-Collados. 2022d. LMMS reloaded: Transformer-based sense embeddings for disambiguation and beyond. *Artificial Intelligence*, 305:103661.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2018. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations.

Elizabeth Closs Traugott. 2017. Semantic change.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. Lscdiscovery: A shared task on semantic change discovery and detection in spanish.

A Sentence-Level Representations

Table 6 presents the comparison results for word-level and sentence-level representations. Since the sentence-level embeddings showed poor performance for the Twitter-RoBERTa-base model (with Macro-F1 of 0.3613), we chose the word-level approach.

	Macro-F1	
Model	**Sentence-level**	**Word-level**
BERT-base-uncased	56.33	57.39
Twitter-RoBERTa-base	36.13	58.26

Table 6: Comparison of sentence-level and word-level representations. Macro-F1 scores for Validation set.