Making History Count

Gabriele Cerizza

Università degli Studi di Milano
gabriele.cerizza@studenti.unimi.it
https://github.com/gabrielecerizza/information_retrieval_project

1 Introduction

In this report we detail our findings in the study of two tasks related to project number 9 for the Information Retrieval course of University of Milan[1]: detection of shifts in the meaning of words across time (Section 2) and extraction of historical events from text (Section 3).

2 Semantic shifts

The first task is concerned with measuring, in a data-driven way, the diachronic semantic shift or lexical semantic change (LSC) that affected words across time. We proceed to review known approaches to tackle the problem and then we discuss three possible strategies to measure LSC by exploiting word embeddings originated from a corpus of historical documents and from a corpus of contemporary documents.

2.1 Related work

In literature, the attempts to capture semantic shifts by way of word embeddings can be categorized in mainly two families: static embeddings or type-based models, and contextualized embeddings or token-based models.

Static embeddings models Static embeddings represent each word with a single vector. This vector can be considered a summarization of the occurrences of a word in different contexts. Popular static embeddings are obtained from skip-gram with negative sampling (SGNS) [11], GloVe [12] and fastText [1].

Static embeddings can be exploited to detect semantic shifts in a straightforward way. We train two models on corpora of different time periods and then we measure the cosine distance between the embeddings of a given word obtained from the two models [6]. The issue in this direct comparison is that SGNS and the other neural models are stochastic in nature and therefore they produce embeddings that are invariant under rotation. Common solutions include:

[1] https://island.ricerca.di.unimi.it/~alfio/shared/inforet/2020-21/inforet-projects.html

- orthogonal Procrustes method, which aligns the embeddings to the same coordinate axes [6], but may introduce noise in the projections [4];
- initialization of the weights of one model with the weights of the other model, which can be problematic for new words [7,16];
- second-order similarity, which compares the cosine similarity of a word to other words in the two models, rather than directly comparing the embeddings of a word [7,16].

Contextualized embeddings models Recently developed neural models like BERT [3] generate different word embeddings for each different context in which a word appears. The pretrained contextualized embeddings are sometimes employed by the authors as is, without fine-tuning on the historical corpus [13,8].

Contextualized embeddings models start by extracting a vector for each context, usually a sentence or a portion thereof, in which the word appears in the two corpora. Then, they cluster the vectors of the two periods to find the different senses of a word. After that, they measure how the frequency of the senses changed between the time periods to get an estimate of the semantic shift [5,13]. An alternative approach computes the average pairwise distance between each vector of one period and each vector of the other period [8].

2.2 Proposed methods

We explored three methods to detect semantic shifts. Two are based on static embeddings and a third one on contextualized embeddings.

Orthogonal Procrustes method (OP) We started by collecting pretrained static word embeddings. For the historical corpus we used the embeddings provided in [15]. These embeddings were trained with fastText on documents dating from 1860 to 1939 and taken from the Corpus of Historical American English (COHA[2]) for a total of 198M tokens. For the contemporary corpus we used fastText word embeddings trained on Wikipedia and news[3] for a total of 16B tokens and 1M word vectors.

Considering the size of the contemporary vocabulary, as well as the fact that it contained misspelled words, we decided to analyze only the 5000 most frequent words in the contemporary vocabulary, intersected with the historical vocabulary. We also made sure to analyze the "target" words indicated by SemEval-2020 [14] for evaluation purposes (see Subsection 2.3).

Finally, we aligned the vectors with orthogonal Procrustes and computed the cosine distance between the two embeddings of each selected word.

[2] https://corpus.byu.edu/coha/

[3] https://fasttext.cc/docs/en/english-vectors.html

Nearest neighbors method (NN) This approach leverages ideas taken from the second-order similarity techniques used with static embeddings. We used the same historical and contemporary word embeddings described for OP. Likewise, we analyzed the same words selected for OP.

The proposed second-order similarity is computed as follows. For each selected word we take the 15 nearest neighbors in the historical vector space; then we measure the cosine distance between the word and these neighbors in the historical vector space; after that, we measure the cosine distance between the word and the same neighbors, but this time in the contemporary vector space; finally we compute the mean squared error (MSE) between the distances measured in the historical vector space and the distances measured in the contemporary vector space. Then we do the same for the 15 nearest neighbors in the contemporary model. Finally, we take the mean between the MSE measured on the historical vector space neighbors and the MSE measured on the contemporary vector space neighbors.

Given this second-order similarity, we decided to shrink the vocabulary of the two models to their intersection. In this way we guaranteed that each neighbor found in one vector space was also present in the other vector space.

Jensen-Shannon distance method (JSD) This method is based on contextualized embeddings. The corpora were taken from SemEval-2020 [14]. The historical corpus contained shuffled sentences from 1810-1860, while the contemporary corpus contained shuffled sentences from 1960-2010. Both corpora were composed of 6M tokens.

We analyzed the 5000 most frequent words in both corpora, keeping only adjectives, nouns and verbs and filtering stop words, punctuations and tokens containing non-alphabetic characters. We aggregated the embeddings according to lemma and POS tags. Our assumption is that the semantic shift of a word is invariant to declension and conjugation: "cat" and "cats" should be subject to the same semantic shift; likewise for "write" and "wrote". Note that in the previous methods we were not able to distinguish words based on POS, since we only had the embeddings, thus the noun "attack" and the verb "attack" had the same embeddings.

We generated the embeddings from BERT, using the "bert-base-cased"[4] pretrained weights, by feeding the transformer with sentences from the corpora. We fine-tuned the model on the historical corpus for 5 epochs. Note that we used only one model to generate embeddings for both time periods. Since BERT generates different embeddings for different contexts, it suffices that BERT acquired knowledge of the contexts used in both time periods [9]. For the embeddings, we took only the hidden state of the last layer, which is reportedly the layer most related to semantics [8].

Following the approach described in [13], we first employed an autoencoder to reduce the dimensionality of the embeddings from 768 to 20. Then, the dimensionality was further reduced to 10 with UMAP. The combination of autoencoder

[4] https://huggingface.co/bert-base-cased

and UMAP was shown to be effective for clustering in [10]. We clustered the embeddings of a given word with HDBSCAN, which is a variant of DBSCAN with improved robustness and a single hyperparameter [2].

Interpreting the clusters as senses of a word, we measured the frequency of each sense in the two time periods. Finally, we used the Jensen-Shannon distance to measure the similarity between the two distributions.

2.3 Results

In Table 1 we show the 10 words whose semantic shift was most prominent according to the proposed methods. In OP we can see many words whose meaning changed due to technological evolution, like “cd”, “tv” and “bot”. OP and NN also capture shifts of historical or cultural nature, like “isis” and “gay”. We observe that the International System of Units was introduced in 1960, hence a possible explanation for the shift in “km”. The semantic shifts detected by JSD are not as readily decipherable.

In Figure 1 we give an example of how the shift in senses was perceived by the JSD method. We can see, for instance, that the verb “score” assumed a meaning more related to sports as the time passed.

Table 1. Words with the most semantic shift detected by the methods.

OP	NN	JSD
cd	deletion	negro (adj.)
romney	km	people (noun)
km	gay	golf (noun)
diff	diff	shopping (noun)
deletion	outstanding	overall (adj.)
tv	implement	businessman (noun)
template	highlight	switch (verb)
isis	parameter	investor (noun)
bot	red	motor (noun)
highlight	template	user (noun)

The instructions provided for project number 9 state that the evaluation of the results will be carried out by experts in the linguistics and history domains. To this end, detailed results for each method can be perused in our repository.

Furthermore, we evaluated our results against a ground truth for the semantic shift occurred in 37 words provided by SemEval-2020 [14]. In Table 2 we compare the Spearman’s rank-order correlation of our methods with the 5 top systems in the SemEval-2020 shared task. We deduce that our methods are competitive with other state-of-the-art systems.

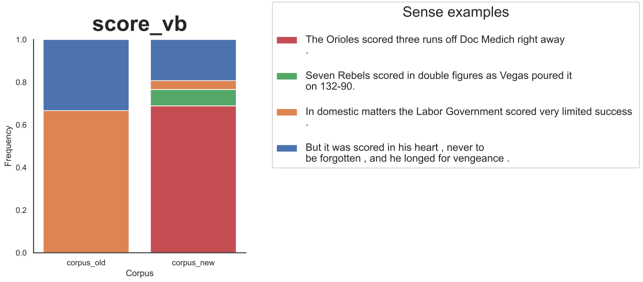

Fig. 1. Shift in the frequency of senses for the verb "score".

Table 2. Spearman's rank-order correlation to SemEval-2020 ground truth for the proposed methods and the top systems in the competition.

OP	NN	JSD	UG_Student_Intern	Jiaxin & Jinan	cs2020	UWB	Discovery_Team
.391	.341	.365	.422	.325	.375	.367	.361

2.4 Conclusion

Tables 1 and 2 suggest that, on the LSC task, static embeddings methods perform slightly better than contextualized embeddings methods. The same conclusion can be found in literature [8,14]. Possible reasons for this behavior are the small size of the corpora for the JSD method and a need of further fine-tuning. Another possible reason is the fact that the historical corpus contains a lot of artifacts and wrongly tokenized words.

Apart from these issues, a possible improvement in the JSD method involves tuning the HDBSCAN hyperparameter to find a number of senses for each word that is similar to the number of word synsets found in WordNet.

3 Historical events extraction

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)
2. Campello, R.J.G.B., Moulavi, D., Sander, J.: Density-based clustering based on hierarchical density estimates. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds.) Advances in Knowledge Discovery and Data Mining. pp. 160–172. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
3. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR (2018), http://arxiv.org/abs/1810.04805

4. Dubossarsky, H., Hengchen, S., Tahmasebi, N., Schlechtweg, D.: Time-out: Temporal referencing for robust modeling of lexical semantic change. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 457–470. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1044, https://aclanthology.org/P19-1044
5. Giulianelli, M., Del Tredici, M., Fernández, R.: Analysing lexical semantic change with contextualised word representations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 3960–3973. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.365, https://aclanthology.org/2020.acl-main.365
6. Hamilton, W.L., Leskovec, J., Jurafsky, D.: 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). pp. 1489–1501. Association for Computational Linguistics (Aug 2016)
7. Kutuzov, A., Øvrelid, L., Szymanski, T., Velldal, E.: Diachronic word embeddings and semantic shifts: a survey. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 1384–1397. Association for Computational Linguistics (Aug 2018)
8. Laicher, S., Kurtyigit, S., Schlechtweg, D., Kuhn, J., im Walde, S.S.: Explaining and improving BERT performance on lexical semantic change detection. CoRR (2021), https://arxiv.org/abs/2103.07259
9. Martinc, M., Kralj Novak, P., Pollak, S.: Leveraging contextual embeddings for detecting diachronic semantic shift. In: Proceedings of the 12th Language Resources and Evaluation Conference. pp. 4811–4819. European Language Resources Association (May 2020)
10. McConville, R., Santos-Rodríguez, R., Piechocki, R.J., Craddock, I.: N2D: (not too) deep clustering via clustering the local manifold of an autoencoded embedding. CoRR (2019), http://arxiv.org/abs/1908.05968
11. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems. vol. 26. Curran Associates, Inc. (2013)
12. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1532–1543. Association for Computational Linguistics (Oct 2014)
13. Rother, D., Haider, T., Eger, S.: CMCE at SemEval-2020 task 1: Clustering on manifolds of contextualized embeddings to detect historical meaning shifts. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. pp. 187–193. International Committee for Computational Linguistics, Barcelona (online) (Dec 2020), https://aclanthology.org/2020.semeval-1.22
14. Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., Tahmasebi, N.: SemEval-2020 task 1: Unsupervised lexical semantic change detection. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation. pp. 1–23. International Committee for Computational Linguistics, Barcelona (online) (Dec 2020), https://aclanthology.org/2020.semeval-1.1
15. Sprugnoli, R., Tonelli, S.: Novel Event Detection and Classification for Historical Texts. Computational Linguistics **45**(2), 229–265 (06 2019)
16. Tahmasebi, N., Borin, L., Jatowt, A.: Survey of computational approaches to diachronic conceptual change. CoRR (2018), http://arxiv.org/abs/1811.06278