DismemBERT: Detecting Diachronic Lexical Semantic Change Using BERT Embeddings

David Rother
TU Darmstadt
david.rother@stud.tu-darmstadt.de

Abstract

This document contains the instructions for preparing a paper submitted to COLING-2020 or accepted for publication in its proceedings. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. These instructions should be used for both papers submitted for review and for final versions of accepted papers. Authors are asked to conform to all the directions reported in this document.

1 Introduction

Here is something regarding semanitc change (Schlechtweg et al., 2018)

2 Related Work

2.1 Diachronic Lexical Semantic Change

With an increasing interest in Diachronic Lexical Semantic Change (LSC) there is a multitude of approaches and three different word representations are commonly used (Schlechtweg et al., 2019).
First are semantic vector representations such as word2vec (Mikolov et al., 2013), which represents each word with two different vectors for each time period respectively (Hamilton et al., 2016a; Hamilton et al., 2016b). The vectors itself represent the co-occurence statistics of the word in the given time period.
Second is the use ditributional representations of words. A word is represented as a vector over all other words occuring in its context. The actual distribution can then be obtained by the word-context co-occurence matrix. To learn these distributions (Frermann and Lapata, 2016) use bayesian learning.
Third are sense clusters where each occurence of a word is assigned to a sense cluster. The clustering usually happens according to some contextual property (Mitra et al., 2015). In newer approaches powerful pretrained deep neural networks such as BERT (Hu et al., 2019; Devlin et al., 2018) are used to extract directly a contextual token of a word from a sentence.

2.2 Word Sense Disambiguation

Word Sense Disambiguation (WSD) is the task of finding different word senses of the same word in sentences. Supervised WSD has sense annotated data and a system usually directly tries to learn Sense Embeddings. The major disadvantage using this approach is that annotating data is very expensive and usually a sufficient amount of data can not be provided.
Unsupervised learning on the other side does not suffer from such constraints. There either leverages some kind of knowledge base such as BabelNet or WordNet, or use a knowledge free model that induces senses (Panchenko et al., 2017).

3 Corpora

The Corpora for evaluation are from the SEMEVAL 2020 Task 1: "Unsupervised Lexical Semantic Change Detection". They contain lemmatized text for english, german, swedish and latin. For each language two corpora are available from two distinct time periods. The respective Time periods can be seen in ref to table. The english corpus is a cleaned version of the COHA corpus (Davies, 2002), where

	t1	t2
English	1810-1860	1960-2010
German	1810-1860	1945-1990
Swedish	1800-1830	1900-1925
Latin	-200-0	0-2000

the corpus has been transformed and every tenth word is replaced by an @. The organizers split the sentences at these tokens and removed them. The german corpus uses the DTA corpus, the BZ ,and the ND corpus.

4 Framework

In this section we present our framework to solve the SemEval 2020 Task 1. We select to apply the same pipeline to all four languages. For each language we have two corpora given and a list of words to compute the LSC on. Since no further fine tuning is done on the provided corpus data we avoid having to align the resulting embedding spaces and it can even be shown that fine tuning may be decreasing with small corpora (Giulianelli, 2019) probably due to overfitting.

4.1 Contextualized Embeddings

We start by computing the contextualized embeddings for each word. To that end we use the Huggingface implementation (Wolf et al., 2019). For the english corpus we compute the embeddings using the bert-base-cased model. The german embeddings are computed with the bert-base-german-cased pretrained model and swedish and latin embeddings are computed using the bert-base-multilingual-cased model that uses the 104 languages with the largest wikipedias.

4.2 Preprocessing

To be able to do efficient clustering on the contextualized embeddings we use a preprocessing pipeline to enhance later results. As a starting point (Reif et al., 2019) show that BERT embeddings projected with UMAP (McInnes et al., 2018), a type of manifold learning similar to t-SNE, produces distinct clusters of the different word senses. Furthermore, (McConville et al., 2019) show that using an autoencoder in conjunction with UMAP leads to higher quality clusterings and that their approach is competetive with other unsupervised deep learning clustering methods. We decide to adopt the latter pipeline and choose a similar autoencoder network with the only change being that we fix the dimension of the latent representation space to be 20 for all words since we have no knowledge of the true amount of senses beforehand.

5 Experiments

In the SemEval-2020 Task 1 challenge there are two different sub-tasks to solve. In the first task one has to decide wether a word has either gained or lost a sense or if the senses remained the same between two time periods. And in the second task the model has to rank words based on the magnitude of change they did undergo. The task organizers quantified the amount of semantic change by constructing a sense frequency distribution for both epochs by human experts. The change score is then the jensen-shannon frequency divergence of the two resulting distributions.

6 Evaluation

The score is ok.

7 Conclusion

This needs work.

References

Mark Davies. 2002. *The Corpus of Historical American English (COHA): 400 million words, 1810-2009*. Brigham Young University.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Lea Frermann and Mirella Lapata. 2016. A bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Mario Giulianelli. 2019. Lexical semantic change analysis with contextualised word representations. *Unpublished masters thesis, University of Amsterdam, Amsterdam*.

William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 2116. NIH Public Access.

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

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.

Ryan McConville, Raul Santos-Rodriguez, Robert J Piechocki, and Ian Craddock. 2019. N2d:(not too) deep clustering via clustering the local manifold of an autoencoded embedding. *arXiv preprint arXiv:1908.05968*.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

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

Sunny Mitra, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, 21(5):773–798.

Alexander Panchenko, Eugen Ruppert, Stefano Faralli, Simone Paolo Ponzetto, and Chris Biemann. 2017. Unsupervised does not mean uninterpretable: The case for word sense induction and disambiguation. Association for Computational Linguistics.

Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B Viegas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of bert. In *Advances in Neural Information Processing Systems*, pages 8592–8600.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (durel): A framework for the annotation of lexical semantic change. *arXiv preprint arXiv:1804.06517*.

Dominik Schlechtweg, Anna Hätty, Marco del Tredici, and Sabine Schulte im Walde. 2019. A wind of change: Detecting and evaluating lexical semantic change across times and domains. *arXiv preprint arXiv:1906.02979*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.