CL-IMS @ DIACR-Ita:
Volente o Nolente: BERT does not Outperform SGNS on Semantic Change Detection

Severin Laicher, Gioia Baldissin, Enrique Castañeda
Dominik Schlechtweg, Sabine Schulte im Walde
Institute for Natural Language Processing, University of Stuttgart
{laichesn,baldisga,medinaeo,schlecdk,schulte}@ims.uni-stuttgart.de*

Abstract

We present the results of our participation in the DIACR-Ita shared task on lexical semantic change detection for Italian. We exploit Average Pairwise Distance of token-based BERT embeddings between time points and rank 5 (of 8) in the official ranking with an accuracy of .72. While we tune parameters on the English data set of SemEval-2020 Task 1 and reach high performance, this does not translate to the Italian DIACR-Ita data set. Our results show that we do not manage to find robust ways to exploit BERT embeddings in lexical semantic change detection.

1 Introduction

Lexical Semantic Change (LSC) Detection has drawn increasing attention in the past years (Kutuzov et al., 2018; Tahmasebi et al., 2018). Recently, SemEval-2020 Task 1 provided a multi-lingual evaluation framework to compare the variety of proposed model architectures (Schlechtweg et al., 2020). The DIACR-Ita shared task extends parts of this framework to Italian by providing an Italian data set for SemEval's binary subtask (Basile et al., 2020a; Basile et al., 2020b). We present the results of our participation in the DIACR-Ita shared task on lexical semantic change for Italian. We exploit Average Pairwise Distance of token-based BERT embeddings (Devlin et al., 2019) between time points and rank 5 (of 8) in the official ranking with an accuracy of .72. While we tune parameters on the English data set of SemEval-2020 Task 1 and reach high performance, this does not transfer to the Italian DIACR-Ita data set. Our results show that we do not manage to find robust ways to exploit BERT embeddings in lexical semantic change detection.

2 Related Work

Most existing approaches for LSC detection are type-based (Schlechtweg et al., 2019; Shoemark et al., 2019). This means that not every word occurrence is considered individually (token-based) but a general vector representation that summarizes every occurrence of a word (including ambiguous words) is created. The results of the SemEval-2020 Task 1 (Martinc et al., 2020; Schlechtweg et al., 2020) showed that type-based approaches (Pražák et al., 2020b; Asgari et al., 2020) achieved better results than token-based approaches (Beck, 2020; Kutuzov and Giulianelli, 2020a). This is somewhat surprising since in the last years contextualized token-based approaches have achieved significant improvements over the static type-based approaches in several NLP tasks (Ethayarajh, 2019). Schlechtweg et al. (2020) suggest a range of possible reasons for this: (i) Contextual embeddings are new and lack proper usage conventions. (ii) They are pre-trained and may thus carry additional, and possibly irrelevant, information. (iii) The context of word uses in the SemEval data set was too narrow (one sentence). (iv) The SemEval corpora were lemmatized, while token-based models usually take the raw sentence as input. In the DIACR-Ita challenge (iii) and (iv) are irrelevant because raw corpora with sufficient context are made available to participants. We tried to tackle (i) by excessively tuning parameters and system modules on the English SemEval data set. (ii) can be tackled by fine-tuning BERT on the target corpora. However, our experiments on the English SemEval data set show that exceptionally high performances can be reached even without fine-tuning.

3 Experimental setup

The DIACR-Ita task definition is taken from SemEval-2020 Task 1 Subtask 1 (binary change detection): Given a list of target words and a diacronic corpus pair C_1 and C_2, the task is to identify the target words which have changed their meanings between the respective time periods t_1 and t_2 (Basile et al., 2020a; Schlechtweg et al., 2020).[1] C_1 and C_2 have been extracted from Italian newspapers and books. Target words which have changed their meaning are labeled with the value '1', the remaining target words are labeled with '0'. Gold data for the 18 target words is semi-automatically generated from Italian online dictionaries. According to the gold data, 6 of the 18 target words are subject to semantic change between t_1 and t_2. This gold data was only made public after the evaluation phase. During the evaluation phase each team was allowed to submit up to 4 predictions for the full list of target words, which were scored using classification accuracy between the predicted labels and the gold data. The final competition ranking compares only the highest of the scores achieved by each team.

4 System Overview

Our model uses BERT to create token vectors and the average pairwise distance to compare the token vectors from two times. The following chapter presents our model, how we have trained it and how we have chosen our submissions.

4.1 BERT

In 2018 Google has released a pre-trained model that ran over Wikipedia and books of different genres (Devlin et al., 2019): BERT (Bidirectional Encoder Representations from Transformer) is a language representation model, designed to find representations for text by analysing its left and right contexts (Devlin et al., 2019). Peters et al. (2018) show that contextual word representations derived from pre-trained bidirectional language models like BERT and ELMo yield significant improvements to the state-of-the-art for a wide range of NLP tasks. BERT can be used to analyse the semantics of individual words, by creating contextualized word representations, vectors that are sensitive to the context in which they appear (Ethayarajh, 2019). BERT can either create one vector for an input sentence (sentence embedding) or one vector for each input token (token embedding).[2]

Different pre-trained BERT models across languages can be downloaded. In this task, we have used the *bert-base-italian-xxl-cased* model for the Italian language[3] to create token embeddings.

The basic BERT version is transformer-based and processes text in 12 different layers. In each layer a contextualized token vector representation can be created for each word in an input sentence. It has been claimed that each layer captures different aspects of the input. Jawahar et al. (2019) suggest that the lower layers capture surface features, the middle layers capture syntactic features and the higher layers capture semantic features of the text. Each layer can serve as representation for the corresponding token by itself, or within a combination of multiple layers.

4.2 Average Pairwise Distance

Given two sets of token vectors from two time periods t_1 and t_2, the idea of Average Pairwise Distance (APD) is to randomly pick a number of vectors from both sets and measure their pair-wise distance (Sagi et al., 2009; Schlechtweg et al., 2018; Giulianelli et al., 2020; Beck, 2020; Kutuzov and Giulianelli, 2020b). The LSC score of the word is the mean average distance of all comparisons:

$$\text{APD}(V, W) = \frac{1}{n_V * n_W} \sum_{v \in V, w \in W} d(v, w)$$

where V and W are two sets of vectors, n_V and n_W denote the number of vectors to be compared, and $d(v, w)$ refer to a distance measure (we used cosine distance (Salton and McGill, 1983)).

4.3 Tuning

The choice of BERT layers and the measure used to compare the resulting vectors (e.g. APD, COS or clustering) strongly influence the performance (Kutuzov and Giulianelli, 2020a). Hence, we tuned these parameters/modules on the English SemEval data (Schlechtweg et al., 2020). For the 40 English

---

[1]The time periods t_1 and t_2 were not disclosed to participants.

target words we had access to the sentences that were used for the human annotation (in contrast to task participants who had only access to the lemmatized larger corpora containing more target word uses than just the annotated ones).

We tested several change measures regarding their ability to find the actual changing words. As part of our tuning, the APD measure produced the binary and graded LSC scores that best matched the actual LSC scores. We also tested the token vectors from different layers in order to check which one fits best to our task. The best layer combinations were the average of the last four layers and the average of the first and last layer of BERT. The highest F1-score for the binary subtask was .75 and a Spearman correlation of .65 for the graded subtask. Our results outperformed all official submissions of the shared tasks, of which the best were all type-based.

4.4 Threshold Selection

We created four predicted change rankings for the target words with BERT+APD. By experience and consideration of the shared tasks (Schlechtweg et al., 2020), we assumed that maximum half of all target words are actual words with a change. Therefore we always annotated at most 9 of 18 words with 1. First, we extracted for each target word a maximum of 200 sentences that contain the word in any token form. We limited the number of uses to 200 for computational efficiency reasons. Then, for each occurrence, we extracted and averaged the token vectors of (i) the last four layers of BERT, and (ii) the first and last layer. For our first submission ('Last Four, 7') we labeled those 7 words with '1' that achieved the highest APD scores in layer combination (i). For our second submission ('First + Last, 7') we labeled those 7 words with '1' that achieved the highest APD scores in layer combination (ii). In (i) and (ii) the same 9 words had the highest APD scores. Therefore, in our third submission ('Average, 9') exactly these 9 words were labeled with '1'. And for our last submission (Lemma, Average, 6') we extracted only sentences in which the target words were present in their lemma form. Again we created the token vectors for the two layer combinations of BERT mentioned above. In both mentioned layer combinations the same 6 words had the highest APD scores. Therefore in our last submission exactly these 6 words were labeled with '1' (similar as in submission 1).

5 Results

Table 1 shows the accuracy scores for the different submissions. The best result was achieved by combining the first and last layer of BERT ('First + Last, 7' with .72), just like on the SemEval data. The second-best result was obtained by using the sentences where the target word occurred in its lemma form ('Lemma, Average, 6' with .67). Only these two submissions outperformed the task baselines and the majority class baseline. The two lowest results were achieved by combining the last four layers of BERT ('Last Four, 7' with .61) and by averaging the two layer combinations ('Average, 9' with .61). The accuracy of our best submission (.72) was ranked at position 5 of the shared task, where the best task result was achieved by two different submissions and reached an accuracy of .94. Both submissions were based on type-based embeddings (Pražák et al., 2020a; Kaiser et al., 2020), clearly outperforming our system.

Submission	Thresh.	Acc.
First + Last	7	**.72**
Lemma, Average	6	.67
Majority Class Baseline	-	.66
Average	9	.61
Last Four	7	.61
Collocations Baseline	-	.61
Frequency Baseline	-	.61

Table 1: Overview accuracy scores for the four submissions with official task baselines. We also report a majority class baseline of a classifier predicting '0' for all target Words.

6 Analysis

As aforementioned, the best performance of our system, achieved with 'First + Last, 7', has an accuracy of .72. It erroneously predicts a meaning change for *cappuccio*, *unico* and *campionato*, while for *palmare* and *rampante* it does not detect the change as given by the gold standard.

We compared both corpora in order to find out if the target words are correctly labeled by the gold standard as well as to identify the possible reasons behind the wrong predictions of our model.

According to our analysis, we can state that the data matches the gold standard. *Cappuccio* is polysemous across both time periods t_0 and t_1 ("hood", "cap"). However, 31% of the uses in t_1 are upper-

cased, namely proper nouns (in contrast to the 4% in t_0), which might imply a different sense compared to the above-mentioned ones:

(1) BENEVENTO Il desiderio di il potere , il potere di il desiderio : ruota intorno a questo inquietante (e attualissimo) spunto il Festival di Benevento diretto da Ruggero **Cappuccio** .

*'BENEVENTO The desire of the power, the power of the desire: the Festival di Benevento directed by Ruggero **Cappuccio** revolves around this unsettling (and current) cue.'*

This skewed distribution of proper names in the two corpora is a possible reason for the wrong prediction of our model.

Throughout all target words, we noticed that the context provided by the previous and the following sentences (as given as input to our model) is often not related topic-wise; in some instances it seems as if the sentences are headlines, since they refer to different topics:

(2) M ROMA Sono quindici gli articoli in cui è suddiviso il provvedimento « antiracket » [...]. Roberta Serra ha vinto ieri lo slalom gigante di il **campionati** italiani femminili .

*'M ROMA The «antiracket» measure is divided into fifteen articles [...]. Roberta Serra won yesterday the giant slalom of the Italian female **championship**.'*

(3) ... le **uniche** azioni pericolose fiorentine sono arrivate quando il pallone e statu giocato su i lati di il Campo . costruzione di centrali idroelettriche , di miniere , canali e strade ...

*'...the **only** dangerous Florentine actions arrived when the ball was played on the sides of the field. Construction of hydroelectric power plants, mines, channels and streets...'*

This "headlines effect" occurs across the whole corpus. It can be traced back to the extraction process of the original corpus and may be a main source of error in our model. Despite not being representative, the following example shows that in some cases no centric window of any size would avoid considering unrelated context.

(4) REPARTO CONFEZIONI UOMO GIACCA cameriere bianca , in tessuto L' **unica** cosa certa è che il governo ha ricevuto una dura lezione da i professori .

*'MEN'S TAILORING DEPARTMENT white textile waiter JACKET The **only** certain thing is that the government has received a hard lesson by the professors.'*

Unico is another example of a word that was erroneously predicted as changing. Due to its abstract meaning ("only", "single", "unique"), it exhibits heterogeneous context across both time periods. Additionally, it can belong to different word classes (noun and adjective in (5) and (6), respectively).

(5) Rischiamo di rimanere gli **unici** a non aver dato mano a la ristrutturazione di le Forze Armate .

*'We risk remaining the **only ones** not having helped in the reorganization of the Armed Forces.'*

(6) ... è chiaro che l' **unica** cosa da fare sarebbe l' unificazione di le due aziende comunali ...

*'...it is clear that the **only** thing to do would be the unification of the two municipal companies...'*

With regards to the undetected changes, the term *palmare* (polysemous within and across word classes) acquires a novel sense in t_1. While it mostly has the meaning of "evident" in the 22 sentences of t_0 (see (7)), it additionally denotes "palmtop" in t_1 (see (8)).

(7) ... con evidenza **palmare** , la impossibilità di difendere una causa perduta ...

*'with **undeniable** evidence, the impossibility of defending a lost cause'*

(8) Per i palestinesi occorre una sistemazione provvisoria in attesa che gli europei si accordino per accoglier li . Potremmo citare in il lungo elenco il **palmare** Apple Newton troppo in anticipo su i tempi

*'A temporary arrangement is needed for the Palestinians while waiting for the Europeans to agree on hosting them. We could quote in the long list the **palmtop** Apple Newton too far ahead of its time'*

Note that also in (8), the topic of the previous and the target sentence is unrelated.

Rampante is a further case of undetected change. The phrase *cavallino rampante*, which metonymically denotes "Ferrari", dominates the usage of the

word in t_0 (70%) and covers a (slightly) relevant share of the uses in t_1 (19%). We hypothesize that this leads to a large number of homogenous usage pairs masking the change from "rampant", "unbridled" to "extremely ambitious" of *rampante*.

7 Conclusion

Our system comprising BERT+APD was ranked 5 in the DIACR-Ita shared task. The combination of BERT and APD did not perform as well as expected and much lower than the best type-based embeddings, but our best submission still outperformed all baselines. The high tuning results achieved on the SemEval data could not be transferred to the Italian data. One reason for this may be that a different BERT model was applied, trained on text of a different language. We have not tuned the Italian BERT model. It is therefore possible that the decrease in performance may be due to the change of the underlying BERT model. Furthermore, given that our model considers as input also the previous and the following sentences, the presence of semantically unrelated context could have played a significant role in mislabeling the target words.

Acknowledgments

References

Ehsaneddin Asgari, Christoph Ringlstetter, and Hinrich Schütze. 2020. EmbLexChange at SemEval-2020 Task 1: Unsupervised Embedding-based Detection of Lexical Semantic Changes. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020a. DIACR-Ita @ EVALITA2020: Overview of the EVALITA 2020 Diachronic Lexical Semantics (DIACR-Ita) Task. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro. 2020b. Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online. CEUR.org.

Christin Beck. 2020. DiaSense at SemEval-2020 Task 1: Modeling sense change via pre-trained BERT embeddings. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

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.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 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 55–65, Hong Kong, China. Association for Computational Linguistics.

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, July. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, July. Association for Computational Linguistics.

Jens Kaiser, Dominik Schlechtweg, and Sabine Schulte im Walde. 2020. OP-IMS @ DIACR-Ita: Back to the Roots: SGNS+OP+CD still rocks Semantic Change Detection. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Andrey Kutuzov and Mario Giulianelli. 2020a. UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Andrey Kutuzov and Mario Giulianelli. 2020b. UiO-UvA at SemEval-2020 Task 1: Contextualised Embeddings for Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. 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. Association for Computational Linguistics.

Matej Martinc, Syrielle Montariol, Elaine Zosa, and Lidia Pivovarova. 2020. Discovery Team at SemEval-2020 Task 1: Context-sensitive Embeddings not Always Better Than Static for Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium, October-November. Association for Computational Linguistics.

Ondřej Pražák, Pavel Přibáň, and Stephen Taylor. 2020a. UWB @ DIACR-Ita: Lexical Semantic Change Detection with CCA and Orthogonal Transformation. In Valerio Basile, Danilo Croce, Maria Di Maro, and Lucia C. Passaro, editors, *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Ondřej Pražák, Pavel Přibáň, Stephen Taylor, and Jakub Sido. 2020b. UWB at SemEval-2020 Task 1: Lexical Semantic Change Detection. In *Proceedings of the 14th International Workshop on Semantic Evaluation*, Barcelona, Spain. Association for Computational Linguistics.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111, Athens, Greece, March. Association for Computational Linguistics.

Gerard Salton and Michael J McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, New York.

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*, pages 169–174, New Orleans, Louisiana, USA.

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 732–746, Florence, Italy. 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.

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*, pages 66–76, Hong Kong, China. Association for Computational Linguistics.

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