CitiusNLP at SemEval-2020 Task 3: Comparing two Approaches for Word Vector Contextualization

Pablo Gamallo
Centro Singular de Investigación en
Tecnoloxías da Información (CiTIUS)
Universidade de Santiago de Compostela, Galiza
pablo.gamallo@usc.es

Abstract

This article describes some unsupervised strategies submitted to SemEval 2020 Task 3, a task which consists of considering the effect of context to compute word similarity. More precisely, given two words in context, the system must predict the degree of similarity of those words considering the context in which they occur, and the system score is compared against human prediction. We compare one approach based on pre-trained BERT models with other strategy relying on static word embeddings and syntactic dependencies. The BERT-based method clearly outperformed the dependency-based strategy.

1 Introduction

The goal of SemEval-2020 Task 3 - Predicting the (Graded) Effect of Context in Word Similarity (Armendariz et al., 2020a), is to predict the degree of similarity of two words considering the context in which those words appear. It should be noted that the context of a word will always have very subtle and continuous (graduated) effects on the lexical meaning, even in the case of those words that are not considered polysemous (Armendariz et al., 2020b). This view of lexical meaning variation as a continuum, inspired in the notion of sense discrimination by Schütze (1998), is quite different from resource-based word sense disambiguation approaches (Navigli and Ponzetto, 2012), which use the local context to predict discrete changes in meaning: the different predefined senses of a polysemous word.

In order to grasp the contextualized sense of words, we will make use of two different strategies:

Transformers: Contextualized word representations produced by BERT, a bi-directional transformer-based language model for context-sensitive word representations (Devlin et al., 2019).

Dependencies: Contextualized word representations produced by considering the static embeddings of syntactically related words to each target word.

This paper is organized as follows. The two strategies are described in Section 2. Experiments and results are presented in Section 3. Finally, conclusions are addressed in Section 4.

2 Two Methods for Word Contextualization

As mentioned in the introduction, we made use of two different strategies, Transformers and Dependencies, to tackle the challenge of contextualizing word senses.

2.1 Transformers and BERT

The first strategy is based on a specific Transformer implementation, namely BERT: Bidirectional Encoder Representations from Transformers (Devlin et al., 2018). Transformers are able to consider word context thanks to the self-attention mechanism, whose objective is to take into account specific parts of the input

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 275–280
Barcelona, Spain (Online), December 12, 2020.

sequence while processing a word. Self-attention allows Transformers to deal with distant contexts of target words, being this mechanism the main difference with regard to Recurrent Neural Networks.

BERT is a bi-directional transformer-based language model that can either be used to extract high quality language features from input text or can be fine-tuned on specific NLP tasks such as entity recognition, classification, question answering, sentiment analysis, and so on. In the present work, we use BERT to extract some features, namely word embedding vectors, from text so as to carry out the graded similarity task of SemEval 2020.

Each transformer layer of 12-layer BERT model stands for a contextualized representation of a given word by putting the focus (or attention) on different chunks of the input sequence. In order to elaborate the individual vectors of each word in context, we combine some of the 12 layers of the deep neural network with the aim of finding the combination of layers that provides the best contextualization of each word in the sequence. By considering that the upper layers of contextualizing word models produce more context-specific representations (Ethayarajh, 2019) which are better suited to the purpose of the task at stake, we create contextualized vectors by combining the last four layers in different ways. In the experiments (next Section), we will show what turned out to be the best layer combination.

For those cases where the BERT tokenizer separates a word into different sub-words (or affixes), we propose to compute the vector average, i.e., to add the contextualized vectors of each sub-word and divide the total by the number of sub-words.

2.2 Dependency-Based Contextualization

According to Rogers et al. (2020), as far as how syntactic information is represented, it seems that syntactic structure is not directly encoded in weights generated by the self-attention mechanism. In fact, the predictions by BERT are not altered even by the recurrent presence of syntactic problems in the input text, such as truncated sentences, shuffled word order, or removed subjects and objects (Ettinger, 2020). This suggests that the BERT's successful encoding of syntactic structure, as in Goldberg's work (2019), does not indicate that it actually relies on that knowledge (Rogers et al., 2020).

Therefore, the second strategy that we are going to carry out is based on the direct use of syntactic dependencies. More precisely, the contextual sense of a given word within a sequence is resulting from considering the static embeddings of those words that are syntactically related with the target word. The Dependency-Based Conceptualized Embedding (DCE) of word w is defined as follows:

$$\text{DCE}(w) = \frac{\sum_{w_i \in DEP_w}^{N} Vec(w_i)}{N} \quad (1)$$

Where DEP_w is the set of words which are either direct dependents or direct heads of w in the input sentence, including w itself; N is the cardinality of this set; and $Vec(w_i)$ is the static vector of word w_i belonging to DEP_w. So, DCE for word w is computed on the basis of those words syntactically related to w by making use of a very simple compositional operation: vector addition. Unlike traditional sentence representation, based on the addition of all words composing the sentence (Mitchell and Lapata, 2008), DCE only pay attention on the most syntactically relevant words of the sentence with regard to the target word. This approach is inspired in previous work for sentence similarity (Gamallo and Pereira-Fariña, 2018), where the similarity between two sentences was not computed by considering all words, but just on the basis of lexical words contained in the argument structure (main verb and their arguments) of the two compared sentences.

3 Experiments

3.1 Task and Subtasks

The datasets provided by the organizers at SemEval-2020 Task 3 - Predicting the (Graded) Effect of Context in Word Similarity were created for four languages: English, Finnish, Croatian, and Slovenian. Annotators were asked to assign a similarity score to a pair of words occurring within a short excerpt of text containing the two words. The two target words appear in two text excerpts with different meanings, which makes their degree of similarity vary as well. The initial test dataset consisted of 340 English word

Word pair	Paragraphs	Rating
manner, *way*	Isaac and Giovanni buried Bento in as Christian **manner** as it was possible under the circumstances, and went to Beijing. After being debriefed by Ricci during a month's stay in Beijing, Isaac returned to India via Macau and the Strait of Singapore, not without some more adventures on the **way**	3
way, *manner*	After escaping to Kenya, he fell in love with hip hop in the **way** that it identified issues being faced by the neighborhood, which he was able to identify with in a unique **manner**. Although he lacked any music background or knowledge of its history, he felt that hip hop could provide the easiest and most effective vehicle to express his story and lobby for political change.	7.3

Table 1: An example of the trial dataset: words *manner* and *way* occurring in two paragraphs with different meanings, and ratings given by the human annotators to the similarity of the two words in each paragraph.

pairs, 24 Finnish pairs, 112 Croatian pairs, and 111 Slovene pairs. In the final evaluation, the organizers also added Estonian word pairs.

In the first paragraph of Table 1, the words *manner* and *way* have very different senses as the second one refers to a spatial path while the first keeps its primary sense: mode of acting. By contrast, in the second paragraph the contextual senses of the two words are very similar as they refer to some kind of mode or style of acting. Human annotators rated as 3 (out of 10) the similarity of the two words in the first paragraph, and 7.3 in the second. 10 is the highest score which should be assigned to full synonyms. The difference between these two scores is 4.3. Computing this difference is the objective of the first subtask as we now go on to describe.

Two subtasks were defined by the organizers: subtask 1, called Predicting Changes, and subtask 2, called Predicting Rating (Armendariz et al., 2020a)

Predicting Changes: The objective is to measure how much each pair of words changes their meaning in two different contexts. Evaluation consists of three steps: first, we compute the difference between the scores produced by the systems when the pair is rated within each one of the two contexts. Second, the same difference is measured with the average of the scores produced by the human annotators. Third, Pearson correlation between the two scores is computed. This subtask evaluates how well systems are able to model the effect that context has in human perception of similarity.

Predicting Ratings: The systems must predict the absolute similarity rating for each pair in each context. The difference between annotators' judgments and system's scores is measured using Spearman correlation.

3.2 Resources and Configuration

Concerning the BERT-based approach, we used the pytorch interface for BERT from HuggingFace's Transformers library (Wolf et al., 2019), BERT tokenizer, and *bert-base-uncased* as pre-trained BERT model for English and *bert-multi-cased* for the rest of languages. Although we have configured the system for all the languages of the test, we have not been able to obtain consistent results in Finnish. Three configurations were implemented by combining the last layers on the neural network in three different ways:

- Adding up the last 4 layers (4last-layers)
- Adding up the last 2 layers (2last-layers)
- Considering only the last layer (1last-layer)

For the dependency-based approach, the static word embeddings used in the experiments were generated with *word2vec* (Mikolov et al., 2013) from the English Wikipedia (December 2018 dump) containing over 2 billion tokens. As Part-of-Speech (PoS) tags are required in the process of contextualization, embeddings were created from PoS tagged text. For this purpose, we used the PoS tagger of the toolkit LinguaKit (Gamallo et al., 2018). To process and parse the input sentences, we made use of DepPattern, a

Models	English	Croatian	Slovenian
1last_layer	**0.7209**	0.4161	0.6224
2last_layers	0.7146	0.2459	**0.6235**
4last_layers	0.7193	**0.4316**	0.5171

Table 2: Pearson correlation obtained in subtask 1 (Change Prediction) on English, Croatian and Slovenian test datasets, by using the Transformer/BERT strategy.

Models	English	Croatian	Slovenian
1last_layer	0.6829	**0.4956**	0.5220
2last_layers	0.6861	0.4486	0.5272
4last_layers	**0.6872**	0.4733	**0.5379**

Table 3: Spearman correlation obtained in subtask 2 (Rating Prediction) on English, Croatian and Slovenian test datasets, by using the Transformer/BERT strategy.

rule-based and multilingual dependency parser (Gamallo and Garcia, 2018; Gamallo, 2015), also taking part of LinguaKit. This approach was only applied on the English dataset.

Several configurations were implemented by considering different PoS tags to extract contextual words:

- Nouns, verbs, and adjectives (NVA)
- Nouns and verbs (NV)
- Nouns and adjectives (NA)
- Only nouns (N)
- Only verbs (V)
- Only adjectives (A)

3.3 Results

Tables 2 and 3 shows the results obtained by using the BERT strategy on English, Croatian and Slovenian languages. Table 2 depicts the Pearson correlation obtained in subtask 1 (Change Prediction), while Table 3 reports the Spearman correlation obtained in subtask 2 (Rating Prediction). None of the implemented model configurations are clearly better than the others. Concerning subtask 1, for the English dataset, the highest value is reached with the 1last_layer. However, for Slovenian, the best model configuration is 2last_layers, while for Croatian, the best one is 4last_layers. In subtask 2, the three configurations give very similar scores even if the 4last_layers seems to perform slightly better. No clear conclusion can be drawn from these results as to which layers of the network are most significant in the process of semantic contextualization.

The results for English in the two subtasks are in the average of those submitted: 8th position (out of 14 systems) in subtask 1 and 8th position (out of 15) in subtask 2. However, for the Slovenian language, the results are relatively much better: 4th position in subtask 1 and 3rd position in subtask 2. It should be noted that in the Slovenian test dataset there are many unknown words that were tokenized by sub-word identification. The strategy chosen to deal with these cases is thus decisive in the final results. The strategy proposed, as has been said, is based on computing the average of adding all sub-words of the target word.

Tables 4 and 5 shows the results obtained by using the dependency-based strategy on only English language. The results are clearly inferior to those obtained with BERT. Even so, it is worth noting the contextual differences that can be seen when comparing different combination of PoS tags. There is a clear difference when the context of a word consists only of nouns or only of adjectives or verbs. Among the main three lexical categories, verbs are the ones that work the worst, while nouns allow the construction of most significant contexts. In fact, single nouns equal or improve results with more complete contexts, such as those composed of nouns and adjectives (NA), nouns and verbs (NV), or nouns, adjectives and verbs (NVA).

Models	English
A	0.1055
V	0.0384
N	0.3325
NA	0.3338
NV	0.2894
NVA	**0.3484**

Table 4: Pearson correlation obtained in subtask 1 (Change Prediction) on English, Croatian and Slovenian test datasets, by using the dependency-based strategy.

Models	English
A	0.4215
V	0.3898
N	0.4243
NA	0.4307
NV	0.3915
NVA	**0.4320**

Table 5: Spearman correlation obtained in subtask 2 (Rating Prediction) on English, Croatian and Slovenian test datasets, by using the dependency-based strategy.

4 Conclusions and Future Work

We have described two different strategies, one based on pre-trained BERT models and other on static word embeddings and syntactic dependencies, to predict the degree of similarity of a pair of words considering the context in which they occur. Evaluation was performed by participating to the SemEval-2020 Task 3: Comparing two Approaches for Word Vector Contextualization.

The BERT-based method clearly outperformed the dependency-based strategy. As for the first one, the four last layers on the neural network were combined in different ways, but no configuration was clearly better than the others. In relation to the syntactic approach, it is worth noting how nouns help to shape syntactic contexts that are more discriminating than those built with only adjectives or just verbs.

In future work, we will try to find a way to experiment with strategies that combine syntactic contexts with Transformer's models.

Acknowledgments

This work has received financial support from DOMINO project (PGC2018-102041-B-I00 , MCIU/AEI/FEDER, UE), eRisk project (RTI2018-093336-B-C21), the Consellería de Cultura, Educación e Ordenación Universitaria (accreditation 2016-2019, ED431G/08, Consolidation and structuring of Groups with Growth Potential: ED431B 2017/39) and the European Regional Development Fund (ERDF).

References

Carlos S. Armendariz, Matthew Purver, Senja Pollak, Nikola Ljubešić, Matej Ulčar, Marko Robnik-Šikonja, Ivan Vulić, and Mohammad Taher Pilehvar. 2020a. SemEval-2020 task 3: Graded word similarity in context (GWSC). In *Proceedings of the 14th International Workshop on Semantic Evaluation*.

Carlos S. Armendariz, Matthew Purver, Matej Ulčar, Senja Pollak, Nikola Ljubešić, and Mark Granroth-Wilding. 2020b. CoSimLex: A resource for evaluating graded word similarity in context. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5878–5886, Marseille, France, May. European Language Resources Association.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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 Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *EMNLP/IJCNLP (1)*, pages 55–65. Association for Computational Linguistics.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suiteof psycholinguistic diagnostics for language models. *TACL*, 8:34–48.

Pablo Gamallo and Marcos Garcia. 2018. Dependency parsing with finite state transducers and compression rules. *Information Processing & Management*, 54(6):1244–1261.

Pablo Gamallo and Martín Pereira-Fariña. 2018. Exploring unsupervised methods to textual similarity. In *Proceedings of the 1st Workshop on Linguistic Tools and Resources for Paraphrasing in Portuguese (POP@PROPOR2018)*, pages 61–87.

P. Gamallo, M. Garcia, C. Piñeiro, R. Martinez-Castaño, and J. C. Pichel. 2018. LinguaKit: A Big Data-Based Multilingual Tool for Linguistic Analysis and Information Extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 239–244.

Pablo Gamallo. 2015. Dependency parsing with compression rules. In *Proceedings of the 14th International Workshop on Parsing Technology (IWPT 2015)*, pages 107–117, Bilbao, Spain. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *CoRR*, abs/1901.05287.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of the Association for Computational Linguistics: Human Language Technologies (ACL-08: HLT)*, pages 236–244, Columbus, Ohio.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193(0):217 – 250.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how BERT works. *CoRR*, abs/2002.12327.

Hinrich Schütze. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–124.

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.