BRUMS at SemEval-2020 Task 3: Contextualised Embeddings for Predicting the (Graded) Effect of Context in Word Similarity

Hansi Hettiarachchi[1], **Tharindu Ranasinghe**[2]
[1]School of Computing and Digital Technology, Birmingham City University, UK
[2]Research Group in Computational Linguistics, University of Wolverhampton, UK
hansi.hettiarachchi@mail.bcu.ac.uk
t.d.ranasinghehettiarachchige@wlv.ac.uk

Abstract

This paper presents the team *BRUMS* submission to SemEval-2020 Task 3: Graded Word Similarity in Context. The system utilises state-of-the-art contextualised word embeddings, which have some task-specific adaptations, including stacked embeddings and average embeddings. Overall, the approach achieves good evaluation scores across all the languages, while maintaining simplicity. Following the final rankings, our approach is ranked within the top 5 solutions of each language while preserving the 1st position of Finnish subtask 2.

1 Introduction

In natural language, meaning of a word has an influence by its surrounding or context words. This impact is mainly driven by the associated linguistic and cognitive phenomena (Armendariz et al., 2020b). Following these two phenomena, it was found that word meanings have a continuous nature in addition to the commonly known discrete nature.

From the cognitive perspective, word meanings can be assigned based on the conceptual structures in human mind (Evans, 2006; Gärdenfors, 2014). Therefore, meaning of a word can be varied according to the mental state of the reader which is triggered with the contact of context words. Thus, word meanings are not just limited to a discrete nature. From the linguistic perspective, context words can modify the meaning of a word by contextual selection and contextual modulation (Cruse et al., 1986). Contextual selection identifies the meanings of polysemous words (i.e. words with multiple senses) such as *'cell'*, *'bank'*, etc. using the context words. In this case, the most appropriate meaning will be picked from a set of discrete senses. For an example, phrase *'prison cell'* implies that the word *'cell'* refers a room. Contextual modulation modifies the meanings of single sense words by highlighting their characteristics. Thus, unlike the contextual selection, modulation makes a continuous effect on the meaning and it is widely available, because majority of the words are general to a certain extent. For an example, *'butter'* is a single sense word. But, if it appears in the phrase *'poured the butter'*, context words reveal that the mentioned butter is in liquid state. Focusing on the above-mentioned facts, it is important to consider both discrete and continuous effects while predicting the meaning of words in natural language text.

Even though there is a continuous effect on meaning, majority of the previous research were only focused on the discrete effect. Considering this limitation, SemEval-2020 Task 3 was designed by targeting the continuous (graded) effects of context. With the involvement of continuous effects, the goal of this task is predicting the effect of context in human perception of similarity. For an example, given the phrases:

> *'Her prison **cell** was almost an improvement over her **room** at the last hostel.'*
>
> *'His job as a biologist didn't leave much **room** for a personal life. He knew much more about human **cells** than about human feelings.'*

this task needs to predict that words *'room'* and *'cell'* have more similar meaning in phrase 1 compared to phrase 2.

Proceedings of the 14th International Workshop on Semantic Evaluation, pages 142–149
Barcelona, Spain (Online), December 12, 2020.

As participants of SemEval-2020 Task 3, we experimented the impact of contextualised word embeddings extracted by different architectures and learning methods on predicting the (graded) effect of context in word similarity. Rather than focusing on the embeddings taken with default settings, we evaluated stacked embeddings, different parameter settings and average embeddings. Further, we experimented the embeddings generated with improved known data rate. In this approach, model unknown words in the data set are replaced with model known words to support more effective embedding generation. The rest of this paper is organised as follows. Section 2 describes the task including its subtasks and data sets. Available methods related to this research are summarised under Section 3 and our approaches are described in Section 4. Following these, results are mentioned in Section 5 and paper is concluded with Section 6.

2 Task description and Data sets

SemEval-2020 Task 3 (Armendariz et al., 2020a) is focused on predicting the (graded) effect of context in word similarity. There were two unsupervised subtasks associated with this shared task as follows;

- **Subtask 1 - Predicting Changes**: Predicting the degree and direction of change in similarity of same word pair within two different contexts (This task targets the ability to identify continuous effects on meaning or effects made by context in human perception of similarity)
- **Subtask 2 - Predicting Ratings**: Predicting the similarity of same word pair within two different contexts (This task is more similar to the traditional task which evaluates the similarity of words based on their contexts)

As the evaluation data set, CoSimLex (Armendariz et al., 2020b) was used. This data set is newly created using human annotators, with the focus on graded effect on word similarity. As the first version, 340 English pairs, 112 Croatian pairs, 111 Slovenian pairs and 24 Finnish pairs were released without human annotated scores to use with the evaluation phase of SemEval-2020 Task 3. Among them, randomly selected 10 English pairs, 5 Croatian pairs and 5 Slovenian pairs were released including the human annotated scores as the practice data for participants to evaluate their models.

3 Related Work

Effect of context was considered by previous research to predict the meaning of words and their similarity. In order to include wider context knowledge, Huang et al. (2012) introduced a neural network-based language model which considers both local and global contexts to learn word representations. As local context, sequence of words is considered and as global context, corresponding document is considered. To capture the multiple senses of a word, they suggested to cluster learned context word vectors into different meaning groups.

In later research, there was a tendency to focus on sense embeddings. Sense embeddings are vectors generated to represent different senses of words. Chen et al. (2014) suggested to use words in WordNet (Miller, 1995) gloss contents to produce sense embeddings. To obtain vector representations of words, they used Skip-gram model (Mikolov et al., 2013). For each polysemous word, a sense was picked by comparing the similarity between its context vector and available sense vectors. Given a sentence, average of its word vectors was computed as the context vector. Similar approach was suggested by another recent research using BERT embeddings (Devlin et al., 2018) to incorporate more contextual features (Loureiro and Jorge, 2019). In this approach, sense embedding generation was initiated using SemCor corpus (Miller et al., 1994). Using the sentences available in the corpus for each sense, BERT embeddings were generated and average of those embeddings were taken as the sense embedding. To improve the sense coverage, WordNet's ontology and glosses also considered. Given a sentence, generated BERT embedding corresponding to a polysemous word was compared with sense embeddings to disambiguate it.

Using meaning groups and sense embeddings, above-mentioned approaches are only looking for discrete word meanings with the effect of context and we could not find any approach which considers

both discrete and continuous effects. As we are aware of CoSimLex is the first data set which is annotated by considering the continuous effect of context and SemEval-2020 Task 3 is the first shared task which is focused on predictions based on this effect. However, most of the available research are based on prediction-based word embeddings and among them contextualised word embeddings are found to be more capable in extracting word meanings based on the context.

4 Methodology

This section describes the different approaches used to predict the (graded) effect of context in word similarity. Since the similarity between a word pair is measured using the cosine similarity between word vectors, we experimented different methods to generate vector representations based on recently published contextualised word embedding models as further described in Sections; 4.1 - 4.4. All the implementations are done in Python [1].

4.1 Contextualised Embeddings

Unlike the classic word embeddings; Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), fastText (Bojanowski et al., 2017), etc., contextualised word embeddings capture the variations of word meanings based on their context. Therefore, they were successfully applied to wide range of NLP tasks such as text classification (Sun et al., 2019; Ranasinghe et al., 2019), question answering (Devlin et al., 2018; Alloatti et al., 2019) and machine translation (Imamura and Sumita, 2019; Zhu et al., 2020) with improved performance.

Among the various contextualised embedding models available, we used ELMo (Peters et al., 2018), Flair (Akbik et al., 2018), BERT (Devlin et al., 2018), Transformer-XL (Dai et al., 2019) and XLNet (Yang et al., 2019) for this task in order to evaluate the impact by different architectures and learning methods on contextual word similarity prediction. Both ELMo and Flair are based on bidirectional LSTM (Long Short-Term Memory) architecture (Sundermeyer et al., 2012) and other models are based on bidirectional Transformer architecture (Devlin et al., 2018). ELMo is learned on sequence of tokens and Flair is learned on sequence of characters. The Transformer architecture introduced by BERT was extended as Transformer-XL by enabling the learning dependencies beyond fixed length contexts. XLNet is a further improved version of Transformer-XL considering the advantages in autoregressive and autoencoding methods. All these Transformer-based models are learned on sequence of tokens.

This research is supported by available pretrained embedding models. Fixed size word vectors are generated for each of the target words using their given context and pretrained model weights. The experimented models including their language coverage are summarised in Table 1. We used the implementations by FLAIR (Akbik et al., 2019)[2] and Hugging Face [3] for embedding generation. More details about the models including layers, parameters and training corpora are available with FLAIR and Hugging Face documentation.

4.2 Stacked Embeddings

Secondly, we tried out the stacked embeddings generated using above-mentioned (Section 4.1) contextualised word embeddings. Stacked embeddings combine each vector by concatenating them to form the final vector (Akbik et al., 2018) as shown in Equation 1. v_i^{stk} represents the final or stacked word vector corresponding to the word i and $v_i^{model_m}$ represents the vector obtained by using the embedding model m. Combining word embeddings from different learning methods allows to combine their characteristics together. For this research, we experimented the stacked embeddings by combining up to three models only.

[1]The code is available on `https://github.com/HHansi/Semeval-2020-Task3`

[2]Git repository of flair is available on `https://github.com/flairNLP/flair`

[3]All the models supported by Hugging Face can be found on `https://huggingface.co/models`

Model	en	hr	sl	fi
ELMo:large	x	-	-	-
Flair:multi-X	x	x	x	x
Flair:hr-X	-	x	-	-
Flair:sl-X	-	-	x	-
Flair:sl-v0-x	-	-	x	-
Flair:fi-X	-	-	-	x
BERT:large-cased	x	-	-	-
BERT:large-uncased	x	-	-	-
BERT:base-multilingual-cased	x	x	x	x
BERT:base-multilingual-uncased	x	x	x	x
BERT:base-finnish-cased-v1	-	-	-	x
BERT:base-finnish-uncased-v1	-	-	-	x
Transformer-XL:wt103	x	-	-	-
XLNet:large-cased	x	-	-	-

Table 1: Embedding models used by this research and their language coverage based on the languages; English(en), Croatian(hr), Slovenian(sl) and Finnish(fi) specific to this task

$$v_i^{stk} = \begin{bmatrix} v_i^{model_1} \\ v_i^{model_2} \\ \vdots \\ v_i^{model_m} \end{bmatrix} \tag{1}$$

4.3 Parameter Settings with BERT

As the next approach, we experimented the effectiveness of BERT embeddings using different parameter settings for embedding extraction layers, sub-token selection and scalar mix. Further, we tested the impact by average embeddings also. BERT model was selected for this experiment, because we could obtain good results for all languages using it.

Embedding extraction layers indicate from which layers of the learned model, weights need to be taken to represent the word vectors. For all the above experiments mentioned in Section 4.1 and 4.2, we used the concatenation of last four layers, because it was found as the best approach to represent the features in underlying text (Devlin et al., 2018).

Since Transformer-based models use sub-tokens to get the embeddings, we analysed the impact by different sub-token selection techniques; first, last, concatenation of first and last (first-last) and mean on predicting the word similarity. Only the first sub-token was used for above experiments.

Scalar mix allows the computation of parameterised scalar mixture of the defined layers (Tenney et al., 2019). This technique was found to be useful, because the best performing layer of a Transformer model could vary depending on the task and it is unclear to the user. Further, Liu et al. (2019) found that scalar mix of Transformer layers has the ability in outperforming the best individual layers. For above experiments, no scalar mix was used.

Average Embeddings: As average embeddings, we considered the average of weights in different layers in order to combine the information learned by them together. For word i, by considering the last k layers, average embedding v_i^{avg} is calculated by following the Equation 2. Weights in the last layer are represented by the vector v_i^{-1}.

$$v_i^{avg} = \frac{v_i^{-k} + ... + v_i^{-1}}{k} \tag{2}$$

4.4 Improved Known Data Rate

Since we used pretrained embedding models for this research, there were tokens such as person names, organisations, locations, etc. in the data set which are unknown to the vocabulary of the pretrained model. By replacing them with some known tokens, we can provide more familiar data to the model during the embedding generation. To convert these tokens into a known form automatically, we used Named Entity Recognition (NER) (Nadeau and Sekine, 2007). Named entities which are identified using the models available with spaCy [4] were used to replace the unknown tokens. For an example, *'...underground in the late 1960s,* ***Sihanouk*** *had to make concessions...'* is converted as *'...underground in the late 1960s,* ***person*** *had to make concessions...'*. But, considering the limited availability of NER models, we conducted this experiment only on English data set.

5 Results

This section contains the results obtained for both subtasks using the approaches described in Section 4. We evaluated the suggested methods using practice data and used the best methods for submissions. The top three scores obtained by submissions on both tasks for all four languages and baseline scores are summarised under Section 5.1 and 5.2.

For all the reported Transformer-based models, as the default parameter setting, concatenation of last four layers using first sub-token without scalar mix is used. Any parameter change except the default setting is mentioned within the brackets after the model name. Stacked models are mentioned using $+$ symbol (e.g. $model_1 + model_2$) and phrase *with NE* is appended to model name if improved known data rate is used.

As the baseline model for both tasks, the multilingual BERT model released by Gary Lai [5] was used as informed by the task organisers. This model was trained using an uncased multilingual data set extracted from Wikipedia.

5.1 Subtask 1

To evaluate the subtask 1 results, Pearson correlation between the model predicted values and gold standards was measured. Average values of scores produced by human annotators are used as gold standards. Since this task is to measure the change in similarity between two contexts, the sign of results is also an important measure, because it indicates the direction of change. Therefore, the uncentered variation of the Pearson correlation which is calculated using the standard deviation from zero was used.

The top three results we obtained with the baseline results for English, Croatian, Slovenian and Finnish are shown in Table 2. According to the results, our approaches could outperform the baseline in all languages except Finnish. English and Croatian used the average embeddings to obtain the best score while Slovenian used the stacked embeddings.

5.2 Subtask 2

To evaluate the subtask 2 results, harmonic mean of Pearson and Spearman correlations between model predicted values and gold standards was used (Camacho-Collados et al., 2017).

The top three resutls we obtained with the baseline results for English, Croatian, Slovenian and Finnish are shown in Table 3. Based on the results, our approaches outperformed the baseline in all languages. The best score in each language was obtained using BERT embeddings which used first-last sub-token with scalar mix.

6 Conclusion

In this paper, we presented different approaches used for SemEval-2020 Task 3: Graded Word Similarity in Context. We mainly evaluated the recent contextualised word embedding models with different embedding generation techniques. Depending on the differences between languages, we could not find any universal approach suitable for all languages to predict the effect of context in word similarity. For subtask 1:

[4] More details about spaCy are available on `https://spacy.io/`

[5] Git repository of baseline embedding model is available on `https://github.com/imgarylai/bert-embedding`

Model	Score
baseline	0.713
BERT:large-cased+Transformer-XL :wt103	0.684
BERT:large-cased(k=14)	**0.754**
BERT:large-cased(k=14):with NE	0.753

(a) Final evaluation results-English

Model	Score
baseline	0.587
BERT:base-multilingual-uncased	0.651
BERT:base-multilingual-cased(first-last)	**0.664**
BERT:base-multilingual-cased(k=4)	**0.664**

(b) Final evaluation results-Croatian

Model	Score
baseline	0.603
Flair:sl-forward+Flair:sl-backward+ BERT:base-multilingual-uncased	**0.648**
BERT:base-multilingual-cased(k=4)	0.608
BERT:base-multilingual-cased(k=6)	0.621

(c) Final evaluation results-Slovenian

Model	Score
baseline	**0.671**
BERT:finnish-cased-v1	0.642
BERT:finnish-uncased-v1	0.594
BERT:finnish-cased-v1(k=6)	0.626

(d) Final evaluation results-Finnish

Table 2: Subtask 1 results on final evaluation

Model	Score
baseline	0.573
BERT:base-multilingual-uncased	0.570
BERT:large-cased(first-last,scalar-mix)	**0.715**
BERT:large-cased(first-last,scalar-mix):with NE	0.713

(a) Final evaluation results-English

Model	Score
baseline	0.402
BERT:base-multilingual-cased	0.482
BERT:base-multilingual-uncased (first-last)	0.528
BERT:base-multilingual-uncased (first-last,scalar-mix)	**0.545**

(b) Final evaluation results-Croatian

Model	Score
baseline	0.516
BERT:base-multilingual-cased	0.524
BERT:base-multilingual-cased(k=4)	0.524
BERT:base-multilingual-uncased(first-last,scalar-mix)	**0.573**

(c) Final evaluation results-Slovenian

Model	Score
baseline	0.289
BERT:finnish-cased-v1	0.644
BERT:finnish-uncased-v1	0.636
BERT:finnish-uncased-v1(first-last, scalar-mix)	**0.645**

(d) Final evaluation results-Finnish

Table 3: Subtask 2 results on final evaluation

predicting changes, best results were obtained using average embeddings and stacked embeddings. It concludes that by combining weights in different layers of same model or different models, degree and direction of change in similarity can be predicted more accurately. For subtask 2: predicting ratings, best results were obtained by BERT embeddings which are generated using first-last sub-token with scalar mix. Different pretrained models were performed well on each language. It concludes that the capability of BERT embeddings on predicting the similarity of words based on their contexts can be further improved using appropriate parameter settings.

As future directions of this research, we hope to experiment the impact by learning and fine tuning options of contextual embedding models on graded effect of context in word similarity.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Francesca Alloatti, Luigi Di Caro, and Gianpiero Sportelli. 2019. Real life application of a question answering system using BERT language model. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 250–253, Stockholm, Sweden, September. Association for Computational Linguistics.

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.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 15–26.

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035.

D Alan Cruse, David Alan Cruse, D A Cruse, and D A Cruse. 1986. *Lexical semantics*. Cambridge university press.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

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*.

Vyvyan Evans. 2006. *Cognitive linguistics*. Edinburgh University Press.

Peter Gärdenfors. 2014. *The geometry of meaning: Semantics based on conceptual spaces*. MIT Press.

Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics.

Kenji Imamura and Eiichiro Sumita. 2019. Recycling a pre-trained BERT encoder for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 23–31, Hong Kong, November. Association for Computational Linguistics.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019. Linguistic knowledge and transferability of contextual representations. 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 1073–1094.

Daniel Loureiro and Alipio Jorge. 2019. Liaad at semdeep-5 challenge: Word-in-context (wic). In *Proceedings of the 5th Workshop on Semantic Deep Learning (SemDeep-5)*, pages 1–5.

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

George A Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the workshop on Human Language Technology*, pages 240–243. Association for Computational Linguistics.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Tharindu Ranasinghe, Marcos Zampieri, and Hansi Hettiarachchi. 2019. Brums at hasoc 2019: Deep learning models for multilingual hate speech and offensive language identification. *In Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation (December 2019)*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In Maosong Sun, Xuanjing Huang, Heng Ji, Zhiyuan Liu, and Yang Liu, editors, *Chinese Computational Linguistics*, pages 194–206, Cham. Springer International Publishing.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. Lstm neural networks for language modeling. In *Thirteenth annual conference of the international speech communication association*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2020. Incorporating bert into neural machine translation. In *International Conference on Learning Representations*.