REPRODUCING THE RESULTS OF THE ICLR 2020 PAPER ”MULTILINGUAL ALIGNMENT OF CONTEXTUAL WORD REPRESENTATIONS”

Alexandru Patrick Pop , Nithish John Paull & Alexander Dennington
School of Electronics and Computer Science
University of Southampton
Southampton, Hampshire, UK
{app1u18, njp1g18,ad2n18}@soton.ac.uk

ABSTRACT

This is a reproduction of the ”Multilingual Alignment of Contextual Word Representations” paper submitted to ICLR 2020. The paper proposes a method of strengthening word alignment to improve the performance of multilingual BERT on downstream tasks. This paper reproduces the results reported initially and attempts to investigate some other factors that might contribute to this process.

1 INTRODUCTION

The research paper that has been attempted to be reproduced[1] in this report is called ”Multilingual Alignment of Contextual Word Representations” (1). This paper aimed to propose novel procedures for evaluating and strengthening contextual word embedding alignment. Thus, validating their utility in analyzing and improving the performance of multilingual BERT (mBERT). The paper includes results that suggest aligning words within parallel corpora before fine-tuning multilingual BERT would lead to improvements, across all languages, on tasks involving zero-shot cross-lingual transfer. The BERT model was trained on the *europarl* (5) dataset, this is a dataset of parallel corpora across 20 different languages. An alignment procedure developed token alignments, where each foreign language word would be mapped to an English word sentence by sentence. The model, after being fine-tuned on the cleaned and aligned version of the dataset, is then integrated within a multi-layer perceptron that is responsible for classifying the natural language inferencing task (NLI). The whole model is then trained on Multi-Genre NLI (MNLI) and tested on the XNLI dataset containing foreign test tasks. These results are of interest because they support contextual alignment as a useful concept for understanding large multilingual pre-trained models.

2 METHODOLOGY

Our aim is to reproduce the contextual alignment process developed in the paper and test to see if we obtain similar results in regards to the testing of the multilingual BERT model on the *europarl* and MNLI data sets. The purpose of this is to check if their methodology is effective generally speaking, i.e. to make sure that there were no special conditions in place for their experiments to succeed.

A model is contextually aligned if it has similar embedding representations of word pairs in parallel sentences. Contextual word retrieval accuracy is measured using the parallel corpus. The contextual word retrieval, mBERT is presented with parallel corpora, where given a word in one language it must find the corresponding word in the other corpora. The paper utilised CSLS (2) to measure the similarity between the embeddings of the word chosen by the model. The model can also pull predictions from target to source corpora - in a reverse direction. Word pairs from the parallel corpora are generated using the *fast_align* function (9). The function can generate word pairs for both source to target pairs and in reverse. The intersection of these two outputs is more valuable, as it removes noise in the word pair alignments and provides an accurate representation of the majority of word pairs within the corpora.

[1] The code is available at https://github.com/COMP6248-Reproducability-Challenge/MULTILINGUAL-ALIGNMENT-OF-CONTEXTUAL-WORD-REPRESENTATIONS/

The fine-tuning step aims to converge the foreign language embeddings with the English language embeddings. By utilising the squared error loss, the similarity between the source and the target embeddings can be quantified and optimised. A regularisation term was defined to maintain the high accuracy of the original model within downstream tasks, like next sentence prediction. The output from the target is compared to the base model's output. This acts as a penalty, keeping the English language embeddings close to their initialisation. Non-English embeddings move closer to their English counterparts after the gradient is derived from the loss function.

The Base mBERT model has trained on parallel corpora from the *europarl* dataset. Since we plan to train the model later to perform NLI as a downstream task, we must pretrain the BERT model using common languages between the *europarl* and XNLI datasets. These languages are Bulgarian, German, Greek, Spanish, and French. The training procedure will only take 250,000 sentences from each corpus to train on. The first 1024 sentences in each corpus will be the testing set, whilst the following will be the validation set. The test set is then altered to include word pairs not present in the training set. The corpora are cleaned of any punctuation or numbers. The paper also takes into account low resource language pairs, where there is a severe problem of data scarcity (3). To combat this the training procedure will learn from 10K and 50K sentences from each corpus, to observe the value of the size of the dataset on the accuracy of prediction.

The finetuned mBERT model is then integrated within a linear classification layer. mBERT utilises the *[CLS]* token to provide embeddings for next sentence prediction. Our model plans to apply a linear layer to predict between three of the NLI classifications, entailment, neutral, or contradiction. The model is given a pair of sentences and predicts whether the first sentence implies the second sentence or not. The combined model is trained on the Multi-Genre NLI corpus (4) and is tested on the XNLI corpus containing non-English translations for NLI. The model is trained for 3 epochs using cross-entropy loss and a linear learning rate warmup scheduler. The development set was used to select the best model, however, due to time and resource constraints, our implementation looks at the first model that was successfully trained for each of the constraints.

3 Analysis

	bg-en	de-en	el-en	es-en	fr-en	Average
Base BERT (Ours)	18.5	28.0	14.9	37.2	35.9	26.9
Base BERT (Original)	19.5	26.1	13.9	32.5	28.3	24.1
Word-aligned BERT (Ours)	43.8	41.6	40.6	46.3	44.8	43.4
Word-aligned BERT (Original)	50.7	51.3	49.8	51.0	48.6	50.3

Table 1: Contextual Word retrieval accuracy for the Base BERT model and the Word-aligned BERT (aligned on 250K sentences)

Sentences	English	Bulgarian	German	Greek	Spanish	French	Average
None	74 (80.4)	62.8 (68.7)	64.4 (70.4)	61.2 (67.0)	68.2 (74.5)	67.3 (73.4)	66.3 (72.4)
10K	75.8 (79.2)	66.3 (71.0)	66.5 (71.8)	62.6 (67.5)	70.1 (75.3)	68.9 (74.1)	68.4 (73.2)
50K	77.8 (81.1)	68.8 (73.0)	68.3 (72.6)	66.3 (69.6)	72.6 (75.0)	71.8 (74.5)	70.9 (74.3)
250K	78.4(80.1)	73.4 (73.4)	71.5 (73.1)	69.6 (71.4)	74.7 (75.5)	73.7 (74.5)	73.5 (74.7)

Table 2: Zero-shot accuracy on the XNLI test set, where we align BERT with varying amounts of parallel data (numbers in brackets represent the original results from the paper).

While the word retrieval accuracy showcased by our model is slightly different from the official published results, we have been able to identify a couple of what we believe are leading causes of this, which we will address later. Nevertheless, we believe that our results showcase that the alignment methodology proposed in the initial paper does work, as it offers an improvement in both contextual word alignment, as well as in the XNLI zero shot task. We will now take a closer look at what might cause the differences in our results from the original reported ones.

For starters, as it is slightly unfeasible to track down the exact version of the Europarl dataset utilised by the authors, we cannot 100% guarantee that we utilised the exact same data. This matters because the testing was done on a relatively small dataset, only 1024 sentences, which means that any deviation in terms of sentences from that will impact the overall accuracy for the Contextual Word Retrieval task, since it changes the values utilised to compute the CSLS distance. For example, changing as little as a couple tens of sentences creates a deviation of around 0.2%. This is a relatively small value by itself, but the effect is still there nonetheless, creating a wider range of deviation as you keep changing the testing corpus, both in size and content.

A secondary factor that we discovered as interesting was the effect of the BERT tokens *[CLS]* and *[SEP]* on the corpus. While we did report the accuracy taking into account those tokens as the main result we obtained, i.e creating CSLS distances for them and attempting to match them to the corresponding tokens in the parallel language, we also wanted to know the effect of removing this attempt at translating them. In essence, when focusing only on the text content of the sentences, all of our BERT models report larger accuracy, as can be seen in Table 3. While these values are indeed slightly better and closer than the originals, we nonetheless feel like excluding tokens was not the intended purpose, as alignment is done with all tokens. This primarily serves as an example of small changes that can and do influence results quite significantly, as well as showcasing that a good chunk of accuracy is lost when attempting to translate tokens. This result we believe is somewhat noteworthy mostly for people in commercial settings, as the main content they are probably interested in translating is represented by the actual words, not the tokens.

A tertiary factor is represented by the pre-trained BERT models chosen. During our experiments, we discovered that pre-trained BERT models will have different embedding values depending on the library they are imported from. We have primarily used the transformers library (7) as it is the newer and officially supported library, however, we also took a small look at the PyTorch-pre trained-Bert (8) library. We also ran some tests on the base multilingual BERT models they provide, measuring the contextual word retrieval accuracy between the two. Here, the same language corpus is fed into both models, and the similarity distances are computed, following the methodology presented in the original paper. The only difference is that instead of trying to map between words in a pair of languages, we try to map words within the same language, from one BERT embedding to the other. Table 4 showcases these accuracies, for all the initial language pairs, alongside the average CSLS distance between them. Note that the values for accuracy and CSLS presented are the average between ”Transformers to Pytorch” and ”Pytorch to Transformers”, however, during our experiments we have discovered that these values do not differ until the 6Th decimal point.

Sentences	Bulgarian	German	Greek	Spanish	French	Average
Base BERT	20.29	30.62	16.91	40.77	38.72	29.46
10K	32.35	33.16	26.76	39.85	38.74	34.17
50K	41.74	37.73	39.92	45.76	42.93	41.61

Table 3: Contextual word retrieval accuracy without tokens

Language	Accuracy contextual word retrieval	Average CSLS
English	88.63	0.014
Bulgarian	79.15	0.014
German	90.76	0.013
Greek	81.90	0.016
Spanish	89.97	0.014
French	92.25	0.013
Average	87.11	0.014

Table 4: Differences in embedding spaces for pretrained multilingual BERT models

4 Reflection & Conclusion

With the contextual word retrieval accuracy being highly correlated with the XNLI Zero-Shot performance in Figure 1, we see that contextual word alignment has a clear impact on the performance of mBERT on downstream tasks. A hypothesis explaining mBERT multilingualism posits that words are forced to have similar representations between languages from common names and numbers (6). Given the veracity of this claim, languages like Bulgarian and Greek that have different scripts than English should experience the worst performance on downstream tasks, we can see this when testing base mBERT. However, after equal

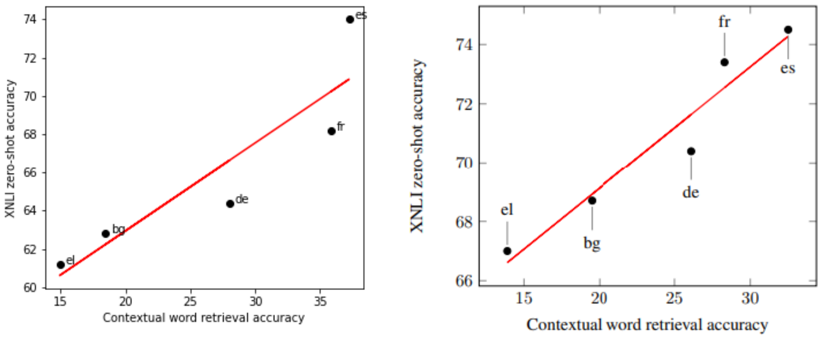

Figure 1: XNLI Zero-Shot Accuracy vs Word Retrieval Accuracies - (Left - Ours) - (Right - Original)

alignment with English, the performance in these languages for both tasks increases dramatically suggesting this training procedure does not represent the contextual alignment within these foreign languages.

In addition to proving their hypothesis, we also experiment with different BERT models as well with slightly changed datasets. These findings show that while this particular alignment procedure is effective at increasing the effectiveness of multilingual BERT on certain tasks, it seems to be dependent on the library utilised for the pretrained model, as well as what the actual data being fed into it.

References

[1] Cao, S., Kitaev, N. and Klein, D., *2020. Multilingual alignment of contextual word representations.* 2020. Multilingual alignment of contextual word representations. arXiv preprint arXiv:2002.03518.

[2] 1. Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herve J'egou. *Word translation without parallel data. In Proceedings of the 6th International Conference on Learning Representations* Word translation without parallel data. In Proceedings of the 6th International Conference on Learning Representations (ICLR 2018), 2018a. URL https://arxiv.org/pdf/1710.04087.pdf

[3] (Dutta Chowdhury et al., 2018) *Multimodal Neural Machine Translation for Low-resource Language Pairs using Synthetic Data* [(https://aclanthology.org/W18-3405)

[4] (Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S.R., Schwenk, H. and Stoyanov, V.,) *XNLI: Evaluating cross-lingual sentence representations* 2018. arXiv preprint arXiv:1809.05053.

[5] Koehn, P., *Europarl: A parallel corpus for statistical machine translation* 2005. In Proceedings of machine translation summit x: papers (pp. 79-86).

[6] Telmo Pires, Eva Schlinger, and Dan Garrette. *How multilingual is multilingual BERT?* InProceedings of the 57th Annual Meeting of the Association for Computational Linguistics,pp. 4996{5001, Florence, Italy, July 2019. Association for Computational Linguistics. URLhttps://www.aclweb.org/anthology/P19-1493.

[7] Transformers library for pretrained BERT: *https://huggingface.co/docs/transformers/index*

[8] pytorch-pretrained-bert library for pretrained BERT: *https://pypi.org/project/pytorch-pretrained-bert/*

[9] fast_align for generating word pairs *https://github.com/clab/fast*$_{a}lign$