Bilingual models for better cross-lingual transfer

Vinícius Jokubauskas

July 2021

Abstract

The use of richer datasets in other languages to fine-tune a network for Natural language processing (NLP) when the target language lacks good datasets has been being explored, and exciting results are being gotten. In this work, we intend to test whether the good performance on the zero-shot tests is due to the similarity of the languages used in the fine-tuning step and the target language or if we can get good results when we use distant language—in lexicology, morphology, and vocabulary. This work evaluates the zero-shot performance in Portuguese for networks pre-trained and several languages (mBert) and Portuguese (BERTimbau) when fine-tuned in Vietnamese. We also investigate the relationship between the sample size in the fine-tuning and the zero-shot precision using several sample sizes from the German, English, and Vietnamese SQuAD based datasets.

1 Introduction

When fine-tuning deep neural networks for specifics tasks, such as the Stanford Question and Answering Dataset (SQuAD) [5] in other languages other than English, the big problem is the size of the size quality of the annotated datasets.

Previous works have shown that Cross-Lingual training fine-tuning a network using a bigger dataset such as SQuAD performs well in another target language, which lacks rich datasets for the fine-tuning [10]. Besides the advantage of getting good results when another well-developed dataset is used to train a network to a specific task in a language with fewer resources for training, there are already studies showing the cost-benefit of such approach [6].

Even though there are exciting results on the cross-lingual performance of networks, such as BERT and mBert [1], studies are showing that the performance is directly correlated with the shared sub-words of the languages used in the fine-tuning and the zero-shot tests [4].

The main goal of this work is to test the zero-shot performance in the Portuguese reading comprehension dataset (FaQuAD) using the Portuguese BERT (BERTimbau) [8] and the Multilingual Bert (mBERT) fine-tuned with the Vietnamese dataset for evaluating machine reading comprehension (ViQuAD) [9], to check if there are any drawbacks in performance when a language with few shared sub-words and other linguistical characteristics are used.

We also are going to test the performance using the truncated versions of SQuAD, GermanQuAD [3] and ViQuAD, to check the correlation between the dataset size and the improvement in F1 Score precision tests.

2 Methodology

2.1 Testing zero-shot with Vietnamese fine-tune

In this section, the metric used is the F1 score as defined in Rajpurkar et al.[5]. Since all datasets tested follows the structure defined in the SQuAD dataset, and the generated answers are based totally on an extracted section of the main corpus, the F1 score is a simple and direct way to evaluate the precision of the models.

There were in total six tests made on this part of the work. Two different networks were used: BERTimbau and mBert. For each of them, we first evaluated the F1 score using the same language in the fine-tuning and the test—in this case, ViQuAD. After that, it was evaluated the zero-shot performance on the Portuguese dataset (FaQuAD) of these networks when trained in Vietnamese. As a baseline for the best F1 possible in both networks, there were made evaluations using only the FaQuAD dataset. The summary of the tests are shown below:

1. BERTimbau fine-tuned using ViQuAD + test using ViQuAD
2. BERTimbau fine-tuned using ViQuAD + test using FaQuAD
3. BERTimbau fine-tuned using FaQuAD + test using FaQuAD
4. mBert fine-tuned using ViQuAD + test using ViQuAD
5. mBert fine-tuned using ViQuAD + test using FaQuAD
6. mBert fine-tuned using FaQuAD + test using FaQuAD

It is important to note that during the training it was used two splits of the ViQuAD dataset, one for training and another for evaluation. It was considered the best F1 Score on the ViQuAD dataset, independently of the language being tested, that way we guarantee that there is no bias in choosing the best F1 Score on the target language datasets.

2.2 Evaluating cross-lingual transfer using truncated datasets

In this section, it was evaluated the F1 Score on the BERTimbau fine-tuned in three different datasets:

1. English (SQuAD)
2. German (GermanQuAD)

3. Vietnamese (ViQuAD)

For each of those languages, the datasets were truncated in 16, 128, 512, 1024, 2048, and 4096 samples. As the previous section, it was considered the best model the one with the best F1 in the validation subset of the language used in the fine-tuning.

3 Data set

The datasets used in all experiments are described below:

FaQuAD: This is a dataset of reading comprehension created by Sayama et al.[7] using data from Brazilian high education institutions. It is based on the SQuAD structure and it consists of 837 questions in the training split and 63 questions in the test split, covering 249 paragraphs taken from official documents.

ViQuad: This is a dataset in Vietnamese for reading comprehension based on the SQuAD structure created by Nguyen et al. [9]. It comprises 23,000 human-generated questions-answer pairs based on 5,109 passages extracted from 174 articles from the Vietnamese Wikipedia.

GermanQuAD: Created by Möller et al. [3], this dataset can be considered the German version of the SQuAD. It consists of 13,722 question-answer pairs based on the German version of Wikipedia.

SQuAD: Probably the most famous dataset for question-answer pairs, it can be considered the most important dataset for question-answer extraction created by Rajpurkar et al. [5] . It is based on more than 100,000 questions, based on 23,215 paragraphs extracted from 536 articles on the English version of Wikipedia.

4 Experiments

In all experiments we used the Colab Tool from Google, all of them performed with an Nvidia Tesla T4 GPU, provided freely by Google. Also, the same hyperparameters and random seeds were used. We used a batch size of 16 and 3 epochs in all tests.

4.1 Testing zero-shot with Vietnamese fine-tune results

During the network's training when we used different languages on the fine-tuning it was noticed that the model tended to over-fit fast, as seen in figure 1.

In Table 2 it is possible to see that there may be a correlation between the dataset size and the performance on the testing, we explore it further in the next subsection. Also, when training the network in Vietnamese, mBert has shown a better F1 Score in the target language, even though it was not pre-trained exclusively in Portuguese. This better performance could be explained

due to the shared vocabulary as explored by Pires et al. [4]. Since mBert was pre-trained using 104 languages, where some of them share characteristics with Vietnamese, being the language itself one of the ones included in the pre-training dataset [2].

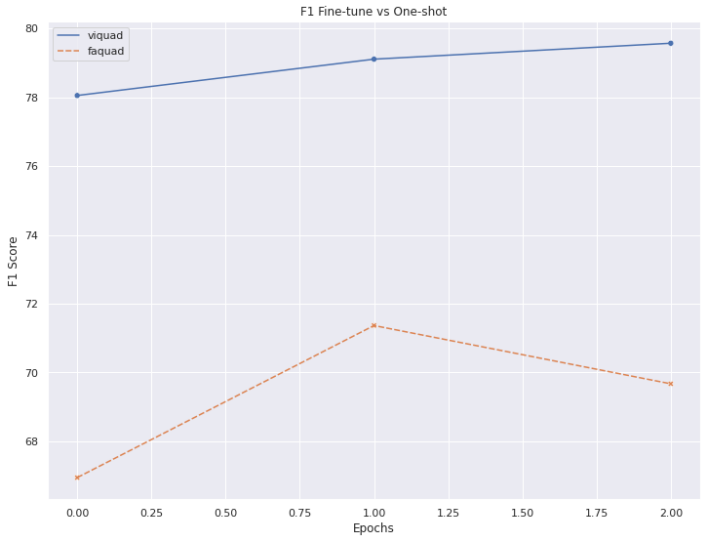

Figure 1: F1 Score for the Validation datase (ViQuAD) and the test dataset (FaQuAD).

Network	**F1 Score**
BERTimbau	63.70
mBert	79.57

Table 1: F1 Score for test and fine-tune made in Vietnamese

4.2 Evaluating cross-lingual transfer using truncated datasets

Since the networks have shown good performance independently of the language used on the fine-tuning, there is the possibility that the size of the datasets to which we train the network may play an important role in the F1 precision on zero-shot tests.

To check the possibility, we evaluated the zero-shot precision using the BERTimbau network trained with a dataset in English (SQuAD), German (GermaQuAD), and Vietnamese (ViQuAD) using several sizes, being the limit a sample size of 4096.

In figure 2 it is possible to see that, with bigger samples sizes, there is an increase in the precision. We also measured the correlation between the sample

Network	Training Dataset	Test Dataset	F1 Score
BERTimbau	(1) FaQuAD	FaQuAD	63.70
	(2) ViQuAD	**FaQuAD**	**43.5**
mBERT	(3) FaQuAD	FaQuAD	64.58
	(4) ViQuAD	**FaQuAD**	**69.67**

Table 2: F1 Score for BERTimbau and mBert for different languages

Sample Size	German	English	Vietnamese
16	15.81	23.19	18.47
128	16.52	17.68	14.55
512	16.00	30.65	12.75
1024	33.83	44.96	17.34
2048	38.45	43.48	23.34
4096	41.34	55.26	33.73

Table 3: F1 Scores in Portuguese (FaQuAD) for several sample sizes

size and the precision for all languages combined, that way it is possible to check if only a difference in the sample size is enough to get good results independently of the language. The correlation measured was 0.74.

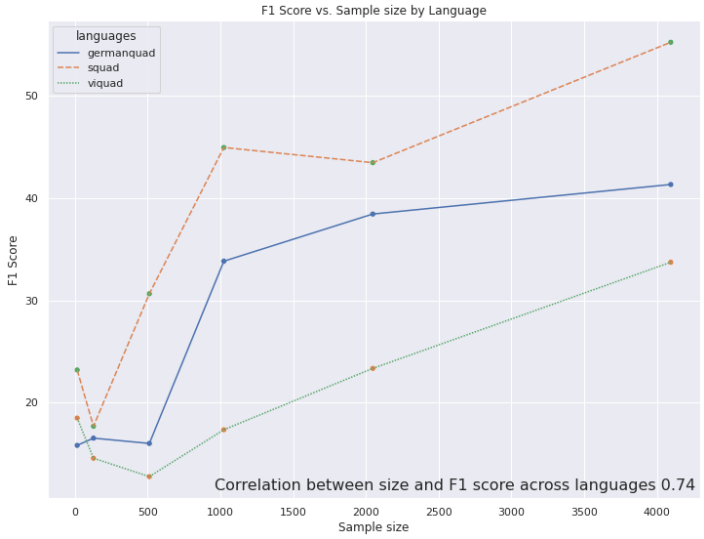

Figure 2: . The F1 Score in Portuguese (FaQuAD) versus the sample size for each language

In Figure 2, it is possible to see that German and English always kept a

higher level of precision if compared with Vietnamese. English and German have also shown higher sensibility with the sample's size increase.

It is well known that German and English share many characteristics and that Vietnamese is a more distant language in morphology and vocabulary. In Figure 3 we compare the precision of each pair of languages and the correlation between the gains in F1 Score.

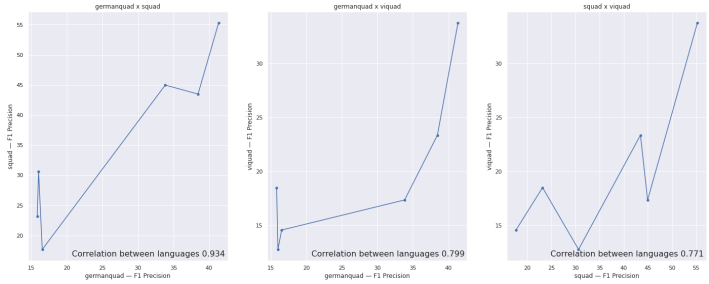

Figure 3: . The correlation in precision between languages

5 Conclusion

This work intended to test the limits of transfer learning when using a language with a distant morphology, lexicology, and vocabulary from the target language. Previous works have shown promising results when similar languages are used for this type of task.

The results in Table 2 explicit that pre-training performed in the language that is going to be used on the fine-tuning step may play an important role in the final task precision, in this case, finding the answer of the question in the corpus body (context).

The results on Table 2 rows (3) and (4) also indicate the importance of the dataset's size. We obtained better results with bigger datasets in a distant language when we used a closer language with smaller datasets in the fine-tuning and the test made in the target language.

This work also investigated the relationship between the F1 Score on zero-shot for different sizes of samples, in Table 3 and Figure 2 it is possible to see that the sample size has its importance on the final precision and that there is a positive correlation between the sizes and the precision, independently of the language that we use in the fine-tuning step.

6 Future Work

The tests relating to the sample size and the precision on the zero-shot evaluation were made using only three languages and tested only in one network

BERTimbau. Also, we used a limited number of sample size variations and the same random seed in all tests.

Further investigation with more languages, sample sizes, and networks could bring stronger statistical results. Also, it could be compared the relationship between the sample size in groups of several similar languages with the target language and in groups of distant languages.

References

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

[2] Jacob Devlin and Slav Petrov. google-research/bert.multilingual.md, Oct 2019.

[3] Timo Möller, Julian Risch, and Malte Pietsch. Germanquad and germandpr: Improving non-english question answering and passage retrieval. *arXiv preprint arXiv:2104.12741*, 2021.

[4] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[5] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[6] Guilherme Moraes Rosa, Luiz Henrique Bonifacio, Leandro Rodrigues de Souza, Roberto Lotufo, and Rodrigo Nogueira. A cost-benefit analysis of cross-lingual transfer methods. *arXiv preprint arXiv:2105.06813*, 2021.

[7] Hélio Fonseca Sayama, Anderson Viçoso Araujo, and Eraldo Rezende Fernandes. Faquad: Reading comprehension dataset in the domain of brazilian higher education. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 443–448, 2019.

[8] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*, 2019.

[9] Kiet Van Nguyen, Duc-Vu Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. A vietnamese dataset for evaluating machine reading comprehension. *arXiv preprint arXiv:2009.14725*, 2020.

[10] Shijie Wu and Mark Dredze. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*, 2019.