Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2020”

Moscow, June 17–20, 2020

EXPLORING THE BERT CROSS-LINGUAL TRANSFER FOR READING COMPREHENSION

Konovalov V. P. (vaskoncv@phystech.edu)[†],
Gulyaev P. A. (guliaev.pa@phystech.edu)[†],
Sorokin A. A. (alexey.sorokin@list.ru)[†, ‡],
Kuratov Y. M. (yurii.kuratov@phystech.edu)[†],
Burtsev M. S. (burtcev.ms@mipt.ru)[†]

[†]Moscow Institute of Physics and Technology, Dolgoprudny, Russia
[‡]Moscow State University, Moscow, Russia

Multilingual BERT has been shown to generalize well in a zero-shot cross-lingual setting. This generalization was measured on POS and NER tasks. We explore the multilingual BERT cross-language transferability on the reading comprehension task. We compare different modes of training of question-answering model for a non-English language using both English and language-specific data. We demonstrate that the model based on multilingual BERT is slightly behind the monolingual BERT-based on Russian data, however, it achieves comparable results with the language-specific variant on Chinese. We also show that training jointly on English data and additional 10,000 monolingual samples allows it to reach the performance comparable to the one trained on monolingual data only.

Keywords: BERT, SQuAD, Question Answering, QA, DeepPavlov

DOI: 10.28995/2075-7182-2020-19-445-453

ИССЛЕДОВАНИЕ КРОСС-ЯЗЫКОВОГО ПЕРЕНОСА МОДЕЛЕЙ ДЛЯ ЗАДАЧИ ОТВЕТА НА ВОПРОСЫ ПО ТЕКСТУ

Коновалов В. П. (vaskoncv@phystech.edu)[†],
Гуляев П. А. (guliaev.pa@phystech.edu)[†],
Сорокин А. А. (alexey.sorokin@list.ru)[†, ‡],
Куратов Ю. М. (yurii.kuratov@phystech.edu)[†],
Бурцев М. С. (burtcev.ms@mipt.ru)[†]

[†]Московский Физико-Технический Институт,
Долгопрудный, Россия
[‡]Московский Государственный Университет
им. М. В. Ломоносова, Москва, Россия

В работе исследуются разные способы выбора данных и модели в задаче обучения вопросно-ответных систем для языка, отличного от английского. Мы показываем, что моноязычный и мультиязычный энкодер приводят к сравнимым результатам при обучении только на данных этого языка. При этом более выгодной стратегией является предварительное обучение мультиязычного энкодера на англоязычных данных с последующей настройкой на данных конкретного языка, поскольку в этом случае требуется значительно меньше обучающих примеров для достижения сопоставимого качества.

Ключевые слова: BERT, SQuAD, Question Answering, QA, DeepPavlov

1. Introduction

Bidirectional Encoder Representations from Transformers (BERT) are the language representation model. Unlike the other models, BERT is designed to pre-train deep bidirectional representations from an unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, including text classification, sequence classification, and question answering, without significant task-specific architecture modifications [4]. Google AI released multiple versions of BERT, such as multilingual BERT (M-BERT), English BERT, and Chinese BERT [6]. M-BERT provides zero-shot cross-lingual model transferability, in which task-specific annotations in one language are used to fine-tune the model for evaluation in another language. The experiments on NER (Named Entity Recognition) and POS (Part-of-speech Tagging) tasks show that while the high lexical overlap between languages improves transfer, M-BERT is also able to transfer between languages having almost no lexical overlap indicating that it captures multilingual representations [10].

In this paper, we measure the cross-lingual model transferability of M-BERT on the reading comprehension task (SQuAD-like datasets) for three languages with small or zero lexical overlap: English, Russian, Chinese. Reading Comprehension is an important task for language understanding, also, it is less susceptible to annotation artifacts found in other datasets [7].

Pre-training BERT for an additional language is a fairly expensive process (four days on 4 to 16 Cloud TPUs) [6]. We show that at least for the Question Answering task this stage is not always necessary, since the obtained performance gain over M-BERT is only marginal, if any. Collecting SQuAD-like datasets takes tremendous efforts and significant funding. Consequently, such datasets rarely exist in languages other than English, making training QA systems in other languages challenging [9]. An alternative to building large monolingual training datasets is to develop cross-lingual systems that can transfer to a target language without requiring training data in that language. We found that M-BERT based model trained jointly on widely available English data and a number of the language-specific monolingual training samples achieves promising results and falls behind the model trained entirely on the language-specific data by a small margin.

Our contributions in this paper are as follows: (i) we show that M-BERT transferability allows us to achieve promising results almost on par with language-specific BERT while both are trained on the language-specific datasets; (ii) even further we show that M-BERT based model jointly trained on English SQuAD with a number of language-specific monolingual samples falls behind the model trained entirely on the language-specific data by a small margin.

We use the **DeepPavlov** framework as our evaluation testbed. **DeepPavlov** is an open-source library for deep learning end-to-end dialog systems and chatbots [2].

2. Related Work

The M-BERT model was introduced in the original BERT paper [4]. It was trained on Wikipedia for 104 languages and uses a shared cross-language BPE vocabulary. Its cross-lingual transferability was demonstrated on sequential tasks like NER and POS. The NER model performance was examined on two publicly available datasets CoNLL-2002 and CoNLL-2003, containing Dutch, Spanish, English, and German. The results showed that the M-BERT performance drops insignificantly when tested on the language that did not participated in the train. The POS experiments were performed by using Universal Dependencies (UD) data for 41 languages. It was shown that M-BERT generalizes well across languages, achieving over 80% accuracy for all pairs. Also, the authors claim that M-BERT's performance is flat for a wide range of overlaps, and even for language pairs with almost no lexical overlap, however, they tested it only on sequential tasks [10].

Moreover, M-BERT was found to be competitive with the state-of-the-art methods for zero-shot cross-lingual transfer on five NLP tasks: natural language inference (XNLI) [3], document classification, NER, POS tagging, and dependency parsing for 39 languages from various language families. Across all five tasks, M-BERT achieved high zero-shot cross-lingual performance without any cross-lingual signal. It outperforms cross-lingual embeddings in four tasks except for XNLI [14].

The previous works came to a consensus that the cross-lingual generalization ability of M-BERT is based on three factors: (i) shared vocabulary items; (ii) joint training across multiple languages (iii) deep cross-lingual representations that generalize across languages and tasks. However, it was shown that the state-of-the-art multilingual representation learning models and a monolingual model that is transferred to new languages at the lexical level perform comparably while tested on standard benchmarks: natural language inference (XNLI dataset), document classification (MLDoc)[12], paraphrase classification (PAWS-X)[15] and question answering (XQuAD)[1]. The authors also showed that a monolingual model trained on a particular language learns some semantic abstractions that are generalizable to other languages[1].

3. Datasets

We measure M-BERT cross-lingual transferability on Reading Comprehension task for three lexical non-overlapping languages: English, Russian, Chinese.

English The Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage (context). SQuAD contains 107,785 question-answer pairs on 536 articles, and is almost two orders of magnitude larger than the previous manually labeled RC dataset [11].

Russia SberQuAD is a Russian reading comprehension dataset analogous to Stanford SQuAD. SberQuAD was collected strictly following the SQuAD's annotations guidelines, which resulted in the high lexical overlap between questions and sentences with answers. We use the same train/test split as in the original paper (45,328/5,036) [5].

Chinese DRCD (Delta Reading Comprehension Dataset) is an open domain traditional Chinese machine reading comprehension dataset. It is based on 2,108 Wikipedia articles [13].

4. Experimental Setup

Our implementation is based on the original model architecture described in [4] and implemented in the `DeepPavlov` framework [2], depicted in **Figure 1**.

We compare QA models' performance based on multilingual BERT (M-BERT) and language-specific BERT (Russian BERT, Chinese BERT). For our models, we use only the `Base` model configuration with 12 hidden layers of 768 units and 12 self-attention heads for input text. The English, Chinese, and multilingual pre-trained BERT were taken from Google's BERT repository. Russian BERT was trained on the Russian part of Wikipedia and news data. Pretraining RuBERT took 250 thousand steps on Tesla P100 x 8 [8]. First, we want to show whether spending expensive computational resources to pre-train language-specific BERT is justified or already available M-BERT can provide us with comparable results. To answer this question we compare language-specific BERT-based models with a multilingual BERT-based model while trained on the same language-specific monolingual training data. We perform comparison in two modes: **`fixed mode`** (**`3 epoch, learning _ rate=5e-05, batch _ size=8, dropout=0.1`**) and **`validated mode`**, where the training step is going until validation patience is achieved. Validation patience (equals to 10) defines how many steps to continue the training process while the performance on the validation set is not improving.

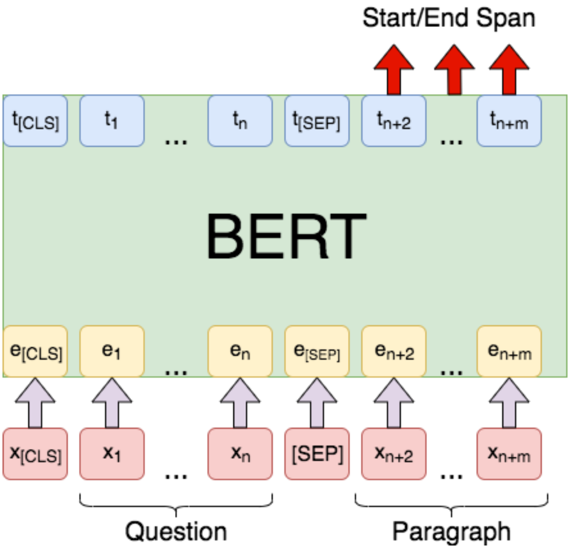

Figure 1: Architecture of the BERT model for SQuAD task

To identify to what extent the language-specific monolingual dataset contributes to the performance of the M-BERT based model, we build learning curves by comparing M-BERT based models in **`validated mode`**.

4.1. Evaluation Metrics

To measure the model's performance we use two metrics Exact Match (EM) and F1-score.

- **Exact match** measures the percentage of predictions that match any one of the ground truth answers exactly.
- **Macro-averaged F1** score measures the average overlap between the prediction and ground truth answers. The prediction and ground truth are bags of tokens.

5. Experimental Results

As expected, validating the models during the training step results in better performance as shown in the Tables.

Table 1: The models' performance (F1/EM) on SQuAD test set (English)

	Fixed mode		Validated mode	
Model Settings (training set)	F1	EM	F1	EM
M-BERT(SQuAD)	80.95	71.68	88.88	81.91
EnBERT(SQuAD)	82.87	74.48	88.67	81.32

The M-BERT based model trained on a language-specific training data substantially outperforms the same model trained on the English training data, this can be observed for both tested languages: **`M-BERT(SberQuAD)`** >> **`M-BERT(SQuAD)`** while tested on the SberQuAD test set and **`M-BERT(DRCD)`** >> **`M-BERT(SQuAD)`** for Chinese. The M-BERT based model trained on English SQuAD fails to extract the exact language-specific answer span that result in significant performance deterioration in EM than in F1.

The RuBERT-based model outperforms the M-BERT based model by a small margin while trained on the SberQuAD training set **`RuBert(SberQuAD)`** > **`M-BERT(SberQuAD)`** in **`validated mode`**, these results are in accordance with the results presented by the authors of RuBERT [8].

Table 2: The models' performance (F1/EM) on SberQuAD test set (Russian)

	Fixed mode		Validated mode	
Model Settings (training set)	F1	EM	F1	EM
M-BERT(SQuAD)	63.5	34.7	73.49	42.94
M-BERT(SberQuAD)	78.78	58.49	83.21	64.56
RuBERT(SberQuAD)	**81.14**	**61.78**	**84.19**	**65.83**

Table 3: The models' performance (F1/EM) on DRCD test set (Chinese)

	Fixed mode		Validated mode	
Model Settings (training set)	F1	EM	F1	EM
M-BERT(SQuAD)	51.73	37.8	75.39	60.28
M-BERT(DRCD)	81.7	75.19	**88.9**	**83.7**
ChBert(DRCD)	**84.46**	**78.01**	88.3	82.67

The DCRD dataset is based on Chinese Wikipedia, it contains Latin symbols that were not properly handled by the character-based tokenization used for pre-training Chinese BERT. This leads to the almost comparable performance of the M-BERT based model with a language-specific Chinese BERT-based model.

Interestingly, the difference in performance between M-BERT and language-specific BERT is much more significant in **`fixed mode`**, that means careful training stopping criteria might mitigate the language-specific BERT superiority.

We build learning curves to measure how the language-specific monolingual dataset contributes to the performance of the M-BERT based model. All models were executed in the **`validated mode`**. The learning curves for Russian are depicted in Figure 2. First, we define two boundaries, the upper bound is the M-BERT based model performance trained solely on the entire SberQuAD dataset (red dashed line). The lower bound is the M-BERT based model performance trained on the English SQuAD dataset (green dashed line). The **`-SQuAD`** curve denotes the model trained on the part of the SberQuAD dataset and the **`+SQuAD`** curve denotes the model trained jointly on the same part of the SberQuAD dataset and on the entire English SQuAD dataset. Adding an entire English SQuAD train to the training set significantly improves performance in comparison to the model trained only on the part of the SberQuAD set. Moreover, the model trained on the joint dataset with 5,000–10,000 language-specific monolingual training samples is only a few points behind the model trained on the entire language-specific training set. The similar findings hold for Chinese as depicted in Figure 3.

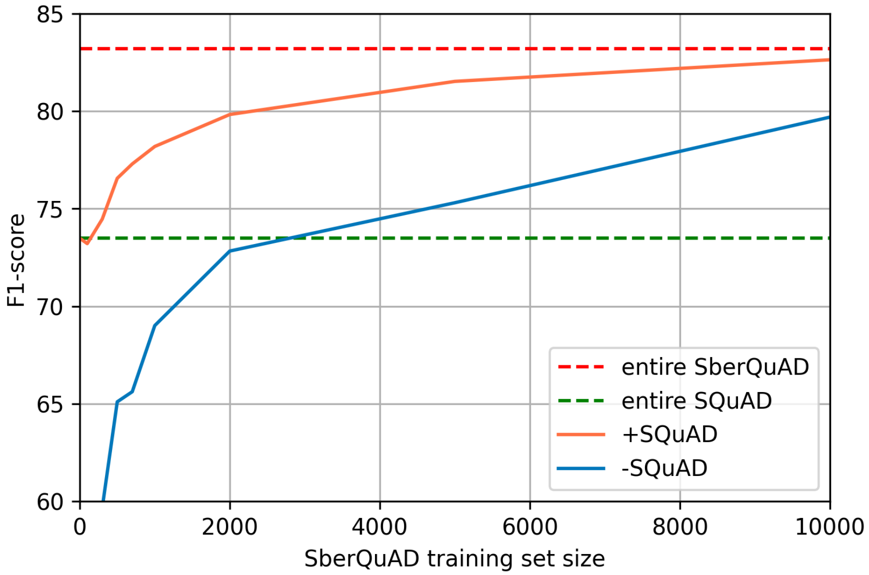

Figure 2: Learning curves on SberQuAD (Russian)

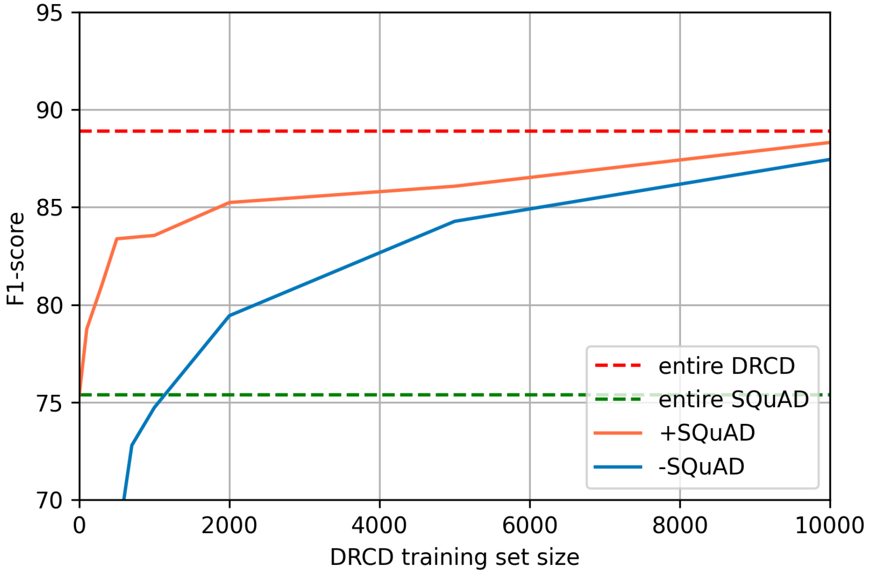

Figure 3: Learning curves on DRCD (Chinese)

6. Conclusion

We measured the multilingual BERT cross-language transferability on the Reading Comprehension task. Specifically, we demonstrated that M-BERT based models perform comparably with language-specific BERT while trained on the same training set. We also showed that M-BERT based models trained jointly on widely available English training data and a number of language-specific instances achieves comparable performance. Our results and analysis agree with the previous theories that M-BERT creates multilingual representations, that allow us to achieve promising performance in cross-lingual model transfer settings. To encourage researchers to further investigate the BERT's cross-language transferability in different tasks we made or code publicly available[1].

7. Acknowledgements

The authors thank all the members of the Deep Learning and Neural Network lab at Moscow Institute of Science and Technology. This work was supported by National Technology Initiative and PAO Sberbank project ID 0000000007417F630002.

[1] http://github.com/deepmipt/DeepPavlov/

References

1. *Artetxe, M. et al.:* On the cross-lingual transferability of monolingual representations. arXiv preprint arXiv:1910.11856. (2019).
2. *Burtsev, M. et al.:* DeepPavlov: An open source library for conversational ai. ACL. (2018).
3. *Conneau, A. et al.:* Xnli: Evaluating cross-lingual sentence representations. arXiv preprint arXiv:1809.05053. (2018).
4. *Devlin, J. et al.:* Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. (2018).
5. *Efimov, P. et al.:* SberQuAD–russian reading comprehension dataset: Description and analysis. arXiv preprint arXiv:1912.09723. (2019).
6. *Google-Research:* TensorFlow code and pre-trained models for bert, (2018).
7. *Kaushik, D., Lipton, Z. C.:* How much reading does reading comprehension require? A critical investigation of popular benchmarks. arXiv preprint arXiv:1808.04926. (2018).
8. *Kuratov, Y., Arkhipov, M.:* Adaptation of deep bidirectional multilingual transformers for russian language. arXiv preprint arXiv:1905.07213. (2019).
9. *Lewis, P. et al.:* Mlqa: Evaluating cross-lingual extractive question answering. arXiv preprint arXiv:1910.07475. (2019).
10. *Pires, T. et al.:* How multilingual is multilingual bert? arXiv preprint arXiv:1906.01502. (2019).
11. *Rajpurkar, P. et al.:* Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250. (2016).
12. *Schwenk, H., Li, X.:* A corpus for multilingual document classification in eight languages. arXiv preprint arXiv:1805.09821. (2018).
13. *Shao, C. C. et al.:* Drcd: A chinese machine reading comprehension dataset. arXiv preprint arXiv:1806.00920. (2018).
14. *Wu, S., Dredze, M.:* Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. arXiv preprint arXiv:1904.09077. (2019).
15. *Yang, Y. et al.:* PAWS-x: A cross-lingual adversarial dataset for paraphrase identification. arXiv preprint arXiv:1908.11828. (2019).