UNIVERSITY OF ZAGREB
FACULTY OF ELECTRICAL ENGINEERING AND COMPUTING

SEMINAR

The Limitations of Zero-Shot Cross-Lingual Transfer

Marko Rajnović
Mentors: *Prof. Dr. Sc. Jan Šnajder; Domagoj Pluščec, Mag. Ing.*

Zagreb, May 2022

CONTENTS

1. Introduction

Zero-shot cross-lingual transfer is the process of pretraining a language model on multiple languages while fine-tuning it on a single one, and evaluating it on some of the pretrained languages. The language of evaluation has to be different from the fine-tuning language. No samples of the language we are evaluating on are present in the fine-tuning dataset. Few-shot cross-lingual transfer has the same pretraining step but differs in fine-tuning. In few-shot transfer, the fine-tuned zero-shot model is additionally fine-tuned on a limited number of samples from the language we will be evaluating on, although fewer than the main language we are using to fine-tune.

In this seminar, we will attempt to replicate some of the results from the paper (Lauscher et al. (2020)). In the paper, the researchers used 2 models, mBERT (Devlin et al. (2019)) and XLM-R (Conneau et al. (2019)). We will focus on the latter, as it is newer and is pre-trained in such a way that it allows for better cross-linguality. The improvement is achieved by the creators of XLM-R sampling their multilingual dataset in a way that the lower-resource languages are oversampled, while the higher-resource languages are undersampled. Multiple transformer models will be fine-tuned on both zero-shot and few-shot datasets and compared, to determine zero-shot effectiveness on a low-level, and a high-level language task. The low-level language task will be NER (named entity recognition) using the WikiAnn dataset (Pan et al. (2017)). The high-level language task will be XNLI (cross-lingual natural language inference) using the XNLI corpus (Conneau et al. (2018)) created by translating dev and test portions of the English Multi-NLI dataset (Williams et al. (2018)). Cross-linguality will be tested on both low-resource and high-resource languages to determine the extent of multilinguality provided by cross-lingual transformer models. Determining zero-shot and few-shot performance quality of language models is important because better performance on these tasks means lower costs for model fine-tuning in different languages and will allow a language product to be language-agnostic.

2. Reproduction report

2.1. Implementation

Although most of the code for the reproduction of the paper was given[1], the reproduction task was far from simple. The main issue we faced in model reproduction was incompatibility issues caused by the lack of a requirements.txt file in the supplied code. The paper was released 2 years before the writing of this seminar, so much has changed with the packages that were used and the code wouldn't run using the current default versions of the libraries used. This was solved mostly through trial and error until the code started functioning, as there was no other way around this problem. The second issue was the VRAM (video random access memory) limitations of the graphics card used for fine-tuning. The fix was to lower the batch size and increase the number of gradient accumulation steps proportionally. The third issue was the sheer number of models that had to be trained for this paper, which put a large strain on the storage of the server these models were trained on. The code for the XNLI task had several errors which were crashing the experiments, most likely due to unavoidable library version differences, so some parts of it had to be rewritten to enable running the experiments. The NER code required more attention, as it was more unfinished. The samplers (k random, k largest, k smallest), that were given for the high-level task, had to be written manually by us. We also had to implement config changes to facilitate proper code reproduction. This was all put together into a bash script to run a large number of models at once. Many issues were caused by the AllenNLP library, as its' older versions have many drawbacks, such as an inability to define when validation is done, as well as difficulties with model evaluation on GPU. The config files were hard to manage, as we never encountered any in this custom format before.

[1] https://aclanthology.org/2020.emnlp-main.363/

2.2. Results

In total, 124 models were trained for the XNLI and NER tasks. For XNLI, 3 were fine-tuned on the large English corpus, and one was chosen to be additionaly fine-tuned in multiple different configurations for 4 different languages, totaling 60 models. A similar setup was done for the NER task.

Task	Model	EN	ΔZH	ΔRU	ΔAR	ΔSW
XNLI	XLM-R	**83.4** (84.3)	-10.2 (-11.0)	-8.68 (-9.0)	-12.65 (-13.0)	-20.42 (-20.2)
NER	XLM-R	**90.2** (91.6)	-34.11 (-34.8)	-9.32 (-13.7)	-41.07 (-24.6)	-19.19 (-20.2)

Table 2.1: Zero-shot cross-lingual transfer performance on XNLI and NER tasks, with the results from the original paper in brackets. We show the accuracy as the change between the english test set and specific language test sets

As shown in Table 2.1,NER doesn't deviate much from the source paper. The performance seems similar enough within a couple of percentage points across both XNLI and NER tasks. The main outlier seems to be Arabic, which performs much worse in our model. The cause of this difference is unknown to us, as we strictly followed the reproduction guidelines set in the paper.

			k=0	k = 10		k=50		k=100		k=500		k=1000	
Task	Model	Sampling	score	score	Δ	score	Δ	score	Δ	score	Δ	score	Δ
XNLI	XLM-R	Random	70.44	70.51	0.07	70.56	0.12	70.31	-0.13	72.92	2.48	73.33	2.89
	XLM-R	Shortest	70.44	70.51	0.07	72.06	1.62	71.26	0.82	73.09	2.65	73.66	3.22
	XLM-R	Longest	70.44	70.53	0.09	70.09	-0.35	70.84	0.4	72.67	2.23	73.08	2.64
NER	XLM-R	Random	64.37	79.28	14.91	83.72	19.35	85.50	21.13	88.63	24.26	89.57	25.2
	XLM-R	Shortest	64.37	61.74	-2.63	60.08	-4.29	61.16	-3.21	65.32	0.95	61.17	-3.2
	XLM-R	Longest	64.37	71.13	6.76	74.63	10.26	74.42	10.05	80.28	15.91	82.69	18.32

Table 2.2: Results of the few-shot experiments with different amounts of target-language examples k, as well as the difference Δ with respect to the zero-shot setting

As shown in Table 2.2, the improvements of few-shot training are largest on the low-level task (NER), while they are substantially smaller on the high-level task (XNLI). This matches the results of the original paper, although our numbers are somewhat different. This will be expanded upon in the next section.

2.3. Conclusion

As we mentioned in Section 2.2, our results are similar to the original paper on the zero-shot task, while they seem to deviate from the original paper in the few-shot experiments. However, this is to be expected, as we used fewer languages due to resource limitations. The NER task benefitted much more from adding k examples, particularly if it was done at random, with just 10 samples seeing an increase of 14.91% accuracy. The performance is worse in the k-shortest and k-longest subtasks, as was in the original paper. The XNLI performance seems to be similar, although slightly worse than in the original paper, but this is to be expected for the above mentioned reason of fewer tested languages.

We could further improve our reproduction analysis in subsequent papers by adding all the languages tested, all the models, as well as the specific grammatical breakdown of languages in the original paper.

3. Bibliography

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, i Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. U *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, stranice 2475–2485, Brussels, Belgium, Listopad-Studeni 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://aclanthology.org/D18-1269.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, i Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, i Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. U *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)*, stranice 4171–4186, Minneapolis, Minnesota, Lipanj 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, i Goran Glavaš. *From Zero to Hero: On the Limitations of Zero-Shot Cross-Lingual Transfer with Multilingual Transformers*, 2020. URL https://arxiv.org/abs/2005.00633.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, i Heng Ji. Cross-lingual name tagging and linking for 282 languages. U *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, stranice 1946–1958, Vancouver, Canada, Srpanj 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1178. URL https://aclanthology.org/P17-1178.

Adina Williams, Nikita Nangia, i Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. U *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, stranice 1112–1122, New Orleans, Louisiana, Lipanj 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101.