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 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 use 2 models, mBERT (Devlin et al. (2019)) and XLM-R (Conneau et al. (2019)), though we will be focusing on the latter, as it is newer and is pre-trained in such a way that it allows for better cross-linguality. The reason for this improvement is in the way that the creators of XLM-R sampled their multilingual dataset, so 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), and the high-level language task will be XNLI (cross-lingual natural language inference). 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 issues

Although most of the code for the reproduction of the paper was given, the reproduction task was far from simple. The main issue we faced in model reproduction were 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 were VRAM limitations of the graphics card used for fine-tuning. The fix was to lower the batch size and increase the number of of gradient accumulation steps proportionally. The third issue was the sheer amount of models that had to be trained for this paper, which put large strain on 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 libary 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 in a more unfinished state. Additional samplers had to be written, as well as config changes to facilitate proper code reproduction. This was all put together into a bash script to run a large amount 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 specific format before.

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**	**ZH**	**RU**	**AR**	**SW**
NER	*X*	**90.2**	-34.11	-9.32	-41.07	-19.19
XNLI	*X*	**83.4**	-10.2	-8.68	-12.65	-20.42

3. Bibliography

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`.