MA-Thesis Research Plan - Pietschke

Daniel Pietschke

November 2024

1 Background and Motivation

With LLMs omnipresent in everyday applications, the question of their impact on our environment has become more and more important to consider. In addition, due to ever increasing technological advancements, people expect continuously better performance from the tools they use. This includes required storage space as well as performance speed and accuracy to the task.

So far the consensus seems to be that bigger models yield better performance due to higher accuracy of their respective tasks. However, bigger models do also mean possible longer processing times as well as an increased CO_2 output every time it is used or trained due to the required processing power and storage space that has to be maintained.

Keeping these impacts in mind, I would like to investigate, how to reduce bigger models effectively and what seems to be the limit in distilling knowledge from a large teacher model, or large teacher models, into a significantly smaller, possibly quicker and more environmentally friendly model. In particular, I would like to investigate the impact this distillation has on the performance of code-switched Machine Translation models. After all, Code-Switching is a prevalent linguistic phenomenon that neural models are struggling with still.

While I am aware of ongoing research into Machine Translation of code-switched language, I am so far unaware that there has been a lot of research into the actual limits of knowledge distillation of these models. Finding a way to effectively reduce needed storage space and retaining accuracy and possibly even increase in performance speed would therefore be a valuable addition to the ongoing research. Even better would it be, if there was some sort of regularity found that can be used in future research into the environmental impact of LLMs and Neural Networks in general.

In addition, I am curious about the effect that multilingual code-switch Machine Translation will have on the end results. Possibly finding an interaction of the limits between bilingual and multilingual models.

2 Research Questions and Goals

Based on my motivation and background that I am aware of for my topic, I will try to answer the question ”What is the limit of Knowledge Distillation for code-switched Machine Translation models?”.

My goal is to use one or several teacher models and train them effectively for the task before using them to train smaller student models and observing the performance development and whether I am able to reduce the number of trainable parameters to the point where performance significantly drops off. I will investigate where this point is and whether there is a regularity to is which can be modeled for future research attempts of similar purposes.

Furthermore, I would like to learn what kind of errors seem to be most common in these smaller models and whether specific information fails to be encoded at some point.

3 Data and Methods

For my thesis, I will use publicly available corpora which have been used in Machine Translation research before or in competitions.

Due to the rarity of proper parallel corpora for code-switched data I will have to use Machine Translation to create the monolingual output data needed for the training of my models. While I am aware, that this compromises the representative results of my research to an extent, I will make sure to use only manually translated corpora as test data to keep the natural representations and I will refrain from using Machine Translation to create code-switched data unless it is absolutely necessary.

Furthermore, I will focus on many-to-one translations for my models as this will facilitate the data collection of manually annotated code-switched translation data.

So far I have found the following datasets:

- Denglish-English dataset: Found in Osmelak and Wintner 2023, 72,801 sentences, but needs Machine Translation
- Spanglish-English dataset: Found in Aguilar, Kar, and Solorio 2020, 21,500 sentences, manually translated
- Hinglish-English dataset: Found in Dhar, Kumar, and Shrivastava 2018, 6,097 sentences, manually translated
- Hinglish-English dataset: Found in Nayak and Joshi 2022, 44,453 sentences, but needs Machine Translation

Looking at this collection it becomes clear that I need much more data if I want to make this research multilingual. Depending on how my data research goes I am considering using the data I have mostly for training and testing whether the model is able to perform somewhat adequately still for unseen languages.

For performing my research, I will use Opus models for the Machine Translation of training data. I will make use of the OpenNMT-py library to build the pipeline for my Knowledge Distillation and creating the student models, aided by the Opus Knowledge Distillation pipeline. For teacher models I am researching unsing pre-trained Opus models for Machine Translation that I can fine-tune for the task at hand. For the actual process of knowledge distillation and finding the limits before performance starts dropping off significantly, I will test different setups for the architecture of the student models, as well as giving different settings to the teacher models' predictions during the training of the student models. Moreover, I will try to find different ways of encoding information for the student models in order to see how different encoding approaches change the outcome. If time permits, I will also try an approach found in Tan et al. 2019 where several bilingual knowledge distillation models are used for training a single multilingual student model. This, however, will depend on how much usable data I will find for training and how much time I will have available for the experiments.

Finally, I will perform statistical analysis on the performance changes of the different student models at different stages of compression. In doing this, I am hoping to uncover the point at which the performance drops off and whether specific settings and changes in the architecture have a stronger impact than others.

4 Preliminary Schedule

I am working still on background research and data collection. I would like to be done with that and have my data ready by the end of December, possibly earlier depending on how the preparation goes.

I would also like to have my models ready and my pipeline for training the teacher models ready by the end of the year/mid-January. After that until Mid-March I would like to create my student models and start the error analysis. After that until the end of May I will write my findings down and create my written thesis to be handed in by the deadline of 22.05.2025.

References

Aguilar, Gustavo, Sudipta Kar, and Thamar Solorio (2020). “LinCE: A centralized benchmark for linguistic code-switching evaluation”. In: *arXiv preprint arXiv:2005.04322*.

Dhar, Mrinal, Vaibhav Kumar, and Manish Shrivastava (2018). “Enabling code-mixed translation: Parallel corpus creation and MT augmentation approach”. In: *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pp. 131–140.

Nayak, Ravindra and Raviraj Joshi (2022). “L3Cube-HingCorpus and HingBERT: A code mixed Hindi-English dataset and BERT language models”. In: *arXiv preprint arXiv:2204.08398*.

Osmelak, Doreen and Shuly Wintner (2023). “The denglisch corpus of German-English code-switching”. In: *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pp. 42–51.

Tan, Xu et al. (2019). “Multilingual neural machine translation with knowledge distillation”. In: *arXiv preprint arXiv:1902.10461*.