Topic: Improve low-resource language performance using intermediate language fine-tuning

Supervisor: Chunlan Ma

Examiner: Prof. Schütze

BSc, MSc: BSc/MSc

Summary:

Multilingual pre-trained language models are widely used today to solve cross-lingual NLP tasks. These models, however, often support only about 100 languages. There are over 7000 languages in the world, most of which lack the possibility of large-scale pre-training or even fine-tuning data. Previous publications have proven intermediate task fine-tuning to be useful in improving target task performance. In this project, we would like to study the possibility of using intermediate languages in the fine-tuning process that enables us to improve the performance of our target language. We will explore, among others, methods of obtaining language similarities and input intermediate languages to different multilingual models in order to compare their task performance.

Prerequisites: good Python programming skills, knowledge of Huggingface and multilingual models, PyTorch

References:

1. What to Pre-Train on? Efficient Intermediate Task Selection (Poth et al., EMNLP 2021)