

Extend the cross-lingual retrieval-augmented prompting method to new tasks and settings

- ☐ **Supervisor:** Ercong Nie
- ☐ **Examiner:** Prof. Hinrich Schütze
- ☐ **BSc/MSc Open:** Msc/Bsc
- ☐ **Summary:** In recent multilingual NLP studies, multilingual pretrained language models (MPLMs), such as mBERT, XLM-R etc., which are jointly pretrained on multilingual corpora, have shown their multilinguality and strong cross-lingual transfer ability[1,2]. However, MPLMs are usually pretrained on unbalanced datasets w.r.t. the distribution of different language resources, and thus MPLMs' representation ability varies greatly among different languages. The main idea of this study is to exploit the MPLMs' strong representation and rich resources of high resource languages (such as English) and to improve the cross-lingual transfer performance for under-represented languages. A method to realize this idea is cross-lingual retrieval-augmented prompting, which is based on the prompt learning[3] method emerging in recent years. Previous study shows that this method takes effect for a set of low-resource languages in classification tasks in zero-shot setup without finetuning[4], indicating potential applications to more fields. This study tries to explore the extensions of this method, for example, to new types of tasks, such as generation tasks (summarization etc.); to other setups, e.g. combined with finetuning; and to new languages that are e.g. not seen in the pretrained data, or even to other modalities than languages. This topic will be co-supervised by Ercong Nie and Sheng Liang.

[1] Shijie Wu and Mark Dredze. 2019. [Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

[2] Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. [From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

[3] Liu, Pengfei, et al. "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing." *arXiv preprint arXiv:2107.13586* (2021).

[4] Nie, Ercong, et al. "Cross-Lingual Retrieval Augmented Prompt for Low-Resource Languages." *arXiv preprint arXiv:2212.09651* (2022).

- ☐ **Prerequisites:** Good programming background in Python; experience in NLP, DL and PyTorch; basic knowledge of transformers; preferably with some linguistic background.