LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
STUDIENGANG COMPUTERLINGUISTIK

Thesis proposal

Topic: **Multilingual Synthetic data generation using LLM**

Supervisor: Hinrich Schütze, Chunlan Ma

Examiner: Hinrich Schütze

Level: MSc

Summary: Data plays a critical role in NLP, serving as the foundation for training and evaluating models. However, the scarcity of high-quality, annotated datasets—especially in the multilingual domain—remains a major obstacle. Languages with fewer speakers are particularly underrepresented in NLP research and applications. Synthetic data generation using Large Language Models (LLMs) provides a scalable and cost-effective solution to this issue. By leveraging the advanced capabilities of LLMs, it is possible to create diverse and linguistically rich datasets that mimic human-generated text. This proposal aims to explore and optimize LLM-based synthetic data generation for multilingual tasks.

- Analyze gaps in existing datasets.
- Use prompts or controlled generation to create realistic synthetic samples.
- Implement quality-check mechanisms.
- Compare the performance of models trained on synthetic data, real data, and their combinations across multiple benchmarks.

Requirements: good programming skills, ability to use large scale language models.

References:

- Yizhong Wang et al. (2023). *Self-Instruct: Aligning Language Models with Self-Generated Instructions*. arXiv: 2212.10560 [cs.CL]. URL: https://arxiv.org/abs/2212.10560
- Abdullatif Köksal et al. (2024). *MURI: High-Quality Instruction Tuning Datasets for Low-Resource Languages via Reverse Instructions*. arXiv: 2409.12958 [cs.CL]. URL: https://arxiv.org/abs/2409.12958
- Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares (Aug. 2024). "Contrasting Linguistic Patterns in Human and LLM-Generated News Text". In: *Artificial Intelligence Review* 57.10. ISSN: 1573-7462. DOI: 10.1007/s10462-024-10903-2. URL: http://dx.doi.org/10.1007/s10462-024-10903-2