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

Thesis proposal

Topic: **Structured Retrieval for Improved Named Entity Recognition with LLMs in historical corpora**

Supervisors: Axel Wisiorek (CIS), Markus Frank (LMU Center for Digital Humanities), Johannes Gleixner (Digitale Geschichtswissenschaft)

Examiner: Axel Wisiorek

Level: BSc/MSc

Keywords: Retrieval-Augmented Generation (RAG), Named Entity Recognition, Entity Linking, Structured Prompting, Factuality, Metadata, Retrieval-Aware Generation

Summary: Entity Linking (EL) in historical and domain-specific texts is complicated by ambiguous entity mentions, spelling variation, and incomplete or noisy coverage of existing knowledge bases. Classical EL pipelines typically follow a NER-first architecture, where entity mentions are detected before being linked to external resources. Recent advances in large language models (LLMs) and Retrieval-Augmented Generation (RAG; Lewis et al. 2020) open up alternative designs in which external knowledge is retrieved and integrated during inference.

In RAG-based pipelines, retrieved documents or knowledge base entries are provided to the LLM as additional context, potentially improving factual grounding and reducing hallucinations. Beyond simple context concatenation, structured prompting and explicit use of retrieval metadata (e.g., document sources, confidence scores, ranks) can further improve factual accuracy, interpretability, and citation behavior.

The goal of this thesis is to study how different ways of integrating retrieval results into LLM-based pipelines affect *named entity recognition (NER) performance*, with *entity linking as an optional downstream task*, while analyzing factuality, robustness, and the impact of structured retrieval integration.

Possible directions and experiments:

- Select or define an evaluation corpus of historical or domain-specific texts with annotated entities.
- Implement a baseline NER pipeline using a standard transformer model or a simple LLM prompt without retrieval.
- Add retrieval of relevant documents or knowledge base entries using dense embeddings (e.g., sentence-transformers) or sparse methods (e.g., BM25/FAISS).
- Integrate retrieved results into the LLM via structured prompts, optionally including metadata such as source, rank, or confidence scores.
- Compare unstructured context concatenation with structured prompting approaches in NER, and optionally assess their impact on downstream EL.
- Evaluate NER performance using standard metrics (F1, precision, recall) and optionally evaluate entity linking accuracy if EL is implemented.
- Conduct error analysis focusing on cases where retrieval helps or hinders predictions, including ambiguous mentions, rare entities, hallucinations, or misattributed links.
- *Optional (advanced):* Encode retrieval metadata as continuous feature vectors and integrate retrieval-aware embeddings (e.g., prefix vectors or virtual tokens) into a frozen LLM, comparing embedding-level integration with prompt-level metadata (cf. Ye et al. 2024).

Resources: The evaluation corpus for this thesis, including historical text annotations as well as external resources such as Wikidata and GND, will be provided by the supervisors.

Requirements: Very good programming and data processing skills in Python, strong background in NLP, familiarity with transformer-based models and large language models, and interest in retrieval-augmented methods and information extraction.

References:

Fan, Wenqi et al. (2024). "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models". In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '24. Barcelona, Spain: Association for Computing Machinery, pp. 6491–6501. ISBN: 9798400704901. DOI: 10.1145/3637528.3671470. URL: https://doi.org/10.1145/3637528.3671470.

Lewis, Patrick et al. (2020). "Retrieval-augmented generation for knowledge-intensive NLP tasks". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. NIPS '20. Vancouver, BC, Canada: Curran Associates Inc. ISBN: 9781713829546.

Li, Yangning et al. (Nov. 2025). "A Survey of RAG-Reasoning Systems in Large Language Models". In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. Ed. by Christos Christodoulopoulos et al. Suzhou, China: Association for Computational Linguistics, pp. 12120–12145. ISBN: 979-8-89176-335-7. DOI: 10.18653/v1/2025.findings-emnlp.648. URL: https://aclanthology.org/2025.findings-emnlp.648/.

Wang, Xiaohua et al. (Nov. 2024). "Searching for Best Practices in Retrieval-Augmented Generation". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 17716–17736. DOI: 10.18653/v1/2024.emnlp-main.981. URL: https://aclanthology.org/2024.emnlp-main.981/.

Ye, Fuda et al. (Nov. 2024). "R^2AG: Incorporating Retrieval Information into Retrieval Augmented Generation". In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 11584–11596. DOI: 10.18653/v1/2024.findings-emnlp.678. URL: https://aclanthology.org/2024.findings-emnlp.678/.