**(Arbeits-)Titel**

Evaluierung von Ansätzen zur Optimierung von Cross Corpus Named Entity Recognition

**Problemstellung**
*Bitte beschreiben Sie Ihre Problemstellung ausführlich und sorgfältig. Eine gut ausformulierte und vorher abgestimmte Beschreibung hilft Ihnen bei der tatsächlichen Bearbeitung Ihrer wissenschaftlichen Arbeit*.

Die automatische Erkennung von Entitäten in Texten (Named Entity Recognition, NER) ist eine zentrale Technologie im Bereich des Natural Language Processing (NLP). Obwohl bestehende NER-Modelle in einem spezifischen Textkorpus häufig hohe Genauigkeitswerte erreichen, zeigen sie erhebliche Leistungseinbußen, wenn sie auf andere Korpora innerhalb derselben Domäne angewendet werden. Diese Abweichungen lassen sich auf Unterschiede in Schreibstilen, Vokabularen, Textlängen, Syntaxvariationen sowie inkonsistente Annotationsrichtlinien zurückführen.

Dies begrenzt die Einsatzfähigkeit von NER-Modellen in realen Anwendungsfällen, bei denen die Daten oft aus verschiedenen Textkorpora stammen.

**Zielsetzung der Arbeit**
*Die Zielsetzung soll möglichst präzise formuliert sein und ggf. auch die Grenzen der Arbeit verdeutlichen. (z.B. Was ist nicht Ziel der Arbeit?)*

Das Ziel der Arbeit ist es die die Leistungsfähigkeit und Generalisierungsfähigkeit von Namen Entity Recognition Modellen in Cross Corpus Umgebungen zu verbessern. Dafür sollen verschiedene Ansätze identifiziert und gegeneinander verprobt werden. Mögliche Ansätze umfassen unter anderem Transfer Learning, Domain Adaptation, Multitask Learning und Datenaugmentation

Der Fokus der Arbeit liegt auf dem Umgang mit unterschiedlichen Textkorpora und nicht auf unterschiedlichen Sprachen oder Domänen. Darüber hinaus soll kein Lösungsansatz für das Problem unterschiedlicher Annotationsrichtlinien gefunden werden, auch wenn dies ein potenzieller Lösungsansatz ist.

**Geplantes Vorgehen**
*Bitte beschreiben Sie Ihre geplante Vorgehensweise detailliert. Dies hilft der wiss. Leitung sehr, die Erfolgswahrscheinlichkeit und Realisierbarkeit Ihrer Arbeit einzuschätzen und Ihnen ggf. wertvolles Feedback zu geben.*

1. Theoretische Grundlagen zu Named Entity Recognition sowie relevanten Evaluationsmetriken
2. Literaturrecherche zu vielversprechenden Ansätzen zur Lösung des Cross Corpus Problems
   2.1. Identifikation und Erklärung relevanter Methoden (Transfer Learning, Domain Adaption, Multitask Learning und Datenaugmentierung, …)
   2.2. Auswahl von vielversprechenden Verfahren, die später evaluiert werden
3. Auswahl geeigneter Korpora sowie deren Bereinigung und Vorverarbeitung
4. Modellentwicklung und Anwendung
   4.1. Entwicklung einer Baseline durch ein einfaches single-corpus Modell
   4.2. Implementierung der ausgewählten Methoden
5. Evaluation der verschiedenen Modelle und Methoden
   5.1. Anwendung der Modelle auf die verschiedenen Korpora und Ermittlung der definierten Evaluationsmetriken
   5.2. Vergleich der Ergebnisse untereinander und mit der Baseline
6. Analyse der Ergebnisse und Ausblick

**Wichtigste Literatur zum Thema**
*Bitte geben Sie die 3-5 wichtigsten Quellen für Ihre Problemstellung an.*

**Devlin J., Chang M-W, Lee K., Toutanova K. (2019):** BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

**Liu, Z., Xu, Y., Yu, T., Dai, W., Ji, Z., Cahyawijaya, S., Madotto, A., & Fung, P. (2021):** CrossNER: Evaluating Cross-Domain Named Entity Recognition. Proceedings of the AAAI Conference on Artificial Intelligence, 35(15), 13452-13460.

**Thomas P, Rocktäschel T, Hakenberg J, Lichtblau Y, Leser U.(2016):** SETH detects and normalizes genetic variants in text, in: Bioinformatics, Volume 32, Issue 18, September 2016, Pages 2883–2885

**Sänger M., Garda S., Wang X., Weber-Genzel L., Droop P., Fuchs B., Akbik A., Leser U. (2024):** HunFlair2 in a cross-corpus evaluation of biomedical named entity recognition and normalization tools; Bioinformatics 2024

**Gururangan S., Marasović A., Swayamdipta S., Lo K., Beltagy I., Downey D., Smith A. (2020):** Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342–8360, Online. Association for Computational Linguistics