Contribution ID: **105** Type: **Talk (15min + 5min)**

NER4all or Context is All You Need: High-Performing Out-of-the-Box NER for Historical and Low-Resource Texts with LLMs through a Humanities-Informed Approach

Wednesday 26 February 2025 11:40 (20 minutes)

Als mit der Veröffentlichung von ChatGPT die Aufmerksamkeit in einem verstärkten Maße auf Large Language Models gelenkt wurde, wusste man zwar, dass sich damit vieles verändern würde, doch die konkreten Auswirkungen waren noch nicht absehbar. Mit unserem Vortrag wollen wir für die Geschichtswissenschaften aufzeigen, was diese neue Technologie ganz konkret für unser Fach leisten kann. Am Beispiel der Named Entity Recognition (NER), also der Erkennung und Klassifikation von Eigennamen bzw. von jenen Textstellen, die namentlich auf Entitäten wie Personen, Orte oder Institutionen verweisen, wollen wir zeigen, inwieweit mit dieser Technologie auch in den Geschichtswissenschaften neue Wege beschritten werden können.

NER ist dabei keineswegs so klar und eindeutig, wie es manchmal den Anschein haben mag –was in besonderer Weise für die Anwendung in den Geschichtswissenschaften gilt. Denn anders als z.B. in der Medizin oder der Biologie, wo diese Techniken in der Regel auf die immer gleichen Textgattungen (zumeist wissenschaftliche Publikationen) angewendet werden, ist die Bandbreite der möglichen Formen und Formate der Quellen in den Geschichtswissenschaften ungleich größer. Sowohl unterschiedliche Sprachen, wie unterschiedliche Textgattungen (von Zeitungsartikeln über Verwaltungsdokumente bis hin zu Briefen oder Tagebucheinträgen), als auch wechselnde Themenbereiche, Sprachebenen wie auch sich historisch wandelnde kulturelle Praktiken und das frühere Fehlen von Rechtschreibregeln bis hin zu unterschiedlichen Ebenen der editorischen Bearbeitung von Quellen bei deren Erschließung erschweren die Aufgabe der Named Entity Recognition erheblich. Zumal hier zuletzt vor allem Methoden ausschlaggebend waren, die auf maschinelles Lernen und damit vor allem auf umfangreich annotierte Trainingsdaten basieren. Für die Geschichtswissenschaften aber ist dies schwierig, da der Aufwand, entsprechende Korpora zu erstellen, äußerst ressourcenintensiv ist und am Ende der Aufwand, eigene Modelle zu trainieren, deren Nutzen übersteigt.

Im Rahmen des Vortrags werden wir zeigen, dass –im Gegensatz zu den bisherigen Behauptungen in der Forschung –LLMs zumindest im Bereich der NER in den Geschichtswissenschaften ein Game Changer sein können. Wir werden zeigen, dass es unter Berücksichtigung der besonderen Eigenschaften der LLM nun möglich ist, ohne den Einsatz spezifischer Modelle oder die aufwändige Aufbereitung von Trainingsdaten und das Training spezifischer Modelle, allein durch geschicktes Prompting, deutlich bessere Ergebnisse zu erzielen als mit den aktuellen Modellen beispielsweise von spaCy und flair. Dabei ist unsere Methode auf jede Textgattung, Domäne und Sprachebene anwendbar.

I want to participate in the youngRSE prize

yes

Primary authors: DRESSELHAUS, Nicole Elisabeth Hitomi (Humboldt-Universität zu Berlin); GRALLERT, Till (Humboldt-Universität zu Berlin); Prof. HILTMANN, Torsten (Humboldt-Universität zu Berlin)

Presenters: DRESSELHAUS, Nicole Elisabeth Hitomi (Humboldt-Universität zu Berlin); GRALLERT, Till (Humboldt-Universität zu Berlin)

Session Classification: Workflows for data pipelines

Track Classification: Data and Software Management: computational workflows