Vorschlag für den Workshop: “Generative KI, LLMs und GPT bei digitalen Editionen”
im Rahmen der Jahrestagung DHd2024

“Zum Einsatz von GPT-4 für NER: Ein Experiment anhand historischer Reisetexte”

Jacob Möhrke, M.A. <moehrke@ios-regensburg.de>*, ORCiD: 0009-0008-6642-5868*
Sandra Balck, M.A. <balck@ios-regensburg.de>*, ORCiD: 0000-0002-0573-3911*
Anna Ananieva, Dr. phil. <ananieva@ios-regensburg.de>*, ORCiD:* 0000-0003-1584-2692
Leibniz–Instituts für Ost- und Südosteuropaforschung (IOS) Regensburg

Der geplante Beitrag setzt sich mit dem Einsatz von KI-basierten Tools im Zusammenhang mit Named Entity Recognition (NER) auseinander und stellt ein Experiment vor, das im Rahmen der digitalen Edition eines Reiseberichts aus dem 19. Jh. (Balck et al. 2023a und 2023b) durchgeführt worden ist. Als Teil des explorativen Vorgehens lotete das Projektteam “Digitale Editionen Historischer Reiseberichte” (DEHisRe) am IOS Regensburg die Anwendungsmöglichkeiten von GPT-4 zwecks NER in einem historischen, als Handschrift überlieferten, deutschsprachigen Reisebericht aus dem Jahr 1810 aus. Im Rahmen dieses Experiments haben wir einen transkribierten und strukturierten Textausschnitt im Umfang von 6.000 Wörtern, der manuell mit 295 Entitäten (Person, Ort, Organisation, Datum, Werk, Währung und Sonstiges) annotiert wurde, mit GPT-4 bearbeitet.

Problemstellung: Historische Texte sind für NER besonders herausfordernd, erstens, weil sie sprachlich von modernen LLMs abweichen, und zweitens, es oft an geeigneten Trainingsdaten mangelt (Ehrmann et al., 2023). Zudem sind LLMs auf Textgenerierung ausgelegt, während NER üblicherweise als Sequenzmarkierungsaufgabe ("sequence labeling task") aufgefasst wird (Wang et al., 2023).

Lösungsansätze: Das Problem der Sequenzmarkierung wurde durch den “GPT-NER”-Ansatz von Wang et al. überwunden, der die NER über angepasste Prompts als Textgenerierungsaufgabe formuliert. Unsere Adaption dieses Ansatzes verwendet spezielle Token (@@, ##, ++, §§, %%, &&, ~~) zur Markierung der Entitäten in der Ausgabe in einem “One-Shot-Learning” Szenario. Die NER wurde im Rahmen unseres Experiments in einem selbst erstellten Custom-GPT durchgeführt, der auf Codeformatierung, Datenkuratierung, TEI-XML und NLP spezialisiert ist.

Linkadressen zum Experiment:

A) Auffinden und Klassifizieren der Entitäten:
https://chat.openai.com/share/79709912-f615-4e12-8c51-f50aeb5c3325
B) Versuch der Normalisierung:
https://chat.openai.com/share/c3a2cf75-15eb-4cfc-a5b1-b05564e4d141
C) Custom-GPT:
https://chat.openai.com/g/g-suHf2Xhom-dh-assistant

Zwischenergebnis: Die Evaluierung gegenüber Handlabeln und anderen Modellen zeigt vielversprechende Ergebnisse. Für die Zwecke der Eruierung und Darstellung der Möglichkeiten ist das GUI gut geeignet, aufgrund der begrenzten Eingabe- und Verarbeitungskapazitäten von GPT wird jedoch eine API-Implementierung für größere Textmengen erforderlich sein. Ein weiteres Problem, das während dem Experiment auftrat, ist die Reproduzierbarkeit der Ergebnisse: In verschiedenen Durchläufen des Experiments erzeugte GPT-4 verschiedene Ergebnisse. Auch hier könnte eine Lösung des Problems in der Verwendung der API liegen, da sie die Modifikation von Parametern wie

“Temperature” ermöglicht. Denn in einem überwachten Umfeld könnte die Varianz der Ergebnisse als Stärke genutzt werden.

Modell	Precision	Recall	F1 Score
GPT-NER	0,86	0,29	0,43
Flair (bisher bestes Modell)	0,69	0,11	0,18

Offene Fragen:

- Die bei dem Experiment demonstrierten operativen Vorteile der KI-Anwendung sprechen für die Eingliederung dieses Tools in den Workflow der digitalen Edition; offen ist allerdings die Frage über den Ort und die Handhabung dieser Anwendung.
- Der generative Character von GPT stellt auch eine Herausforderung für die Implementierung der Editionsrichtlinien dar: das GPT hat u.a. eigenständig Schritte zur Normalisierung unternommen und dabei zur Emendierung der Textquelle beigetragen, die den diplomatischen Charakter der Transkription der historischen Quelle verletzt. Beispiel (Output GPT): "Den 19.ten" (the 19th, referring to the next day after the 18th of February 1810) is labeled as: ~~19. Febr. 1810%%

Literatur:

Sandra Balck, Hermann Beyer-Thoma, Ingo Frank, Anna Ananieva (2023a). “Interlinking Text and Data with Semantic Annotation and Ontology Design Patterns to Analyse Historical Travelogues”. In: Digital Humanities Quarterly 17, 3 (Special Issue “Working on and with Categories for Text Analysis: Challenges and Findings from and for Digital Humanities Practices”), 17 pages. http://www.digitalhumanities.org/dhq/vol/17/3/000726/000726.html (09.01.2023).

Sandra Balck, Jacob Möhrke, Anna Ananieva (2023b). “Digitale Editionen historischer Reiseberichte: Der Kreislauf historischer Informationen”. In: Digital Humanities Day Leipzig 2023 (DHDL 23). Forum für Digital Humanities Leipzig (4. Dezember 2023), Leipzig. https://doi.org/10.5281/zenodo.10301629.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet (2023). “Named Entity Recognition and Classification in Historical Documents: A Survey”. In: ACM Comput. Surv. 56, 2, Article 27 (February 2024), 47 pages. https://doi.org/10.1145/3604931.

Xiaoya Li (2023). “A Practical Survey on Zero-Shot Prompt Design for In-Context Learning”. In: Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing (pp. 641–647). INCOMA Ltd., Shoumen, Bulgaria. https://aclanthology.org/2023.ranlp-1.69.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang (2023). “GPT-NER: Named Entity Recognition via Large Language Models”. https://doi.org/10.48550/arXiv.2304.10428.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, Xinyun Chen (2023). “Large Language Models as Optimizers”. https://doi.org/10.48550/arXiv.2309.03409.