

eScriptorium meets LLMs: Moderne KI-Systeme im Kontext der Volltexterschließung

Will, Larissa

larissa.will@uni-mannheim.de
Universitätsbibliothek Mannheim, Deutschland
ORCID: 0009-0004-6220-8939

Kamlah, Jan

jan.kamlah@uni-mannheim.de
Universitätsbibliothek Mannheim, Deutschland
ORCID: 0000-0002-0417-7562

Schmidt, Thomas

thomas.schmidt@uni-mannheim.de
Universitätsbibliothek Mannheim, Deutschland
ORCID: 0000-0003-3620-3355

Lang, Sarah

sarah.lang@uni-graz.at
Zentrum für Informationsmodellierung, Universität Graz,
Österreich
ORCID: 0000-0002-4618-9481

Huff, Dorothee

dorothee.huff@uni-tuebingen.de
Universitätsbibliothek Tübingen, Deutschland
ORCID: 0000-0003-0866-9967

Einführung

In diesem Workshop werden wir mit der Texterkennungs- und Transkriptionsplattform eScriptorium (Stokes et. al, 2021) historische Handschriften und Drucke automatisiert segmentieren und transkribieren. Anschließend werden die Ergebnisse mit Large Language Models (LLMs) (Zhao et al. 2023, S. 2-7) wie ChatGPT (ChatGPT, 2024) weiterverarbeitet. Moderne Texterkennungssoftware bietet mittlerweile viele Möglichkeiten und erleichtert die Arbeit mit historischen Dokumenten, beispielsweise für digitale Editionen, ungemein.

Jedoch ist es kein Geheimnis, dass diese bei historischen Texten immer noch häufig an ihre Grenzen stößt. Sei es ein kompliziertes Layout wie Tabellen oder bei typographischen und graphematischen Besonderheiten. Vor allem im Kontext der hohen Ansprüche, die an die Korrektheit digi-

taler Editionen gestellt werden, ist der Output nicht ohne (teils aufwändige) Nachbearbeitung verwendbar.

Mit der Einführung von Transformer-Modellen (Wikipedia, 2024) in die Engine Kraken (Kraken, 2024), dem Kern von eScriptorium, erwarten wir einen qualitativen Sprung in der Qualität der automatischen Texterkennung. Aber wie groß ist dieser Sprung wirklich? Und wie können LLM-Systeme die Erstellung von Trainingsmaterial im Kontext eScriptorium sowie die Erschließung der Volltexte für weitere Fragestellungen unterstützen?

Die Rolle von Conformer-Modellen in Kraken

Die Implementierung von Conformer-Modellen (Kiessling, 2024) in die Kraken-Engine stellt einen bedeutenden Fortschritt dar und folgt damit anderen Transkriptionsplattformen wie Transkribus, die mit ihren sogenannten Supermodellen (READ COOP, 2023) bereits eine Transformer-Variante etablieren konnten. Die hybride Transformer-Architektur Conformer (convolution-augmented Transformer) kombiniert die Stärken von Convolutional Neural Networks (CNNs) und Transformer-Architekturen, um die Texterkennung und -verarbeitung weiter zu verbessern. Die CNN-Komponente hilft bei der Erkennung lokaler Merkmale in Bildausschnitten, während die Transformer-Komponente die Erkennung längerer kontextueller Abhängigkeiten in Sequenzen ermöglicht. Als Transformer-Komponente wurde die Transformer-XL-Architektur gewählt, die eine bessere Leistung bei kontextuellen Abhängigkeiten erzielt: 80 % länger als RNNs und 450 % länger als herkömmliche Transformer. Sowohl bei kurzen als auch bei langen Sequenzen wird eine bessere Leistung erzielt und die Auswertung erfolgt bis zu 1800-mal schneller als bei herkömmlichen Transformatoren (Dai et. al, 2019, 2978-2986). Diese neuen Modelle sollen in eScriptorium als "drop in replacements"¹ für die alten Modelle verwendet werden können.

Anwendung von LLMs zur Textverarbeitung und -anreicherung

Large Language Models wie ChatGPT bieten eine Vielzahl von Möglichkeiten zur Weiterverarbeitung und Anreicherung der Ergebnisse aus der Texterkennung. LLMs können beispielsweise genutzt werden, um automatisch Python-Skripte zu generieren, die zur Strukturierung von Textdaten verwendet werden. Darüber hinaus können sie Fehler in der Texterkennung korrigieren und gezielt Informationen aus dem Volltext extrahieren.

Historische Dokumente wie Protokolle, Tagebücher und Briefnachlässe stellen besondere Herausforderungen für die Texterkennung dar. Im Gegensatz zu modernen Druckwerken fehlen häufig Inhaltsverzeichnisse, Register oder

Klappentexte, die inhaltliche Rückschlüsse zulassen. Dies erschwert die gezielte Suche und Erschließung solcher Materialien. In vielen Fällen bleibt nur eine zeitintensive manuelle Verschlagwortung oder Regestierung, um die Zugänglichkeit zu erhöhen. Hier können KI-Methoden einen entscheidenden Beitrag leisten. Durch den Einsatz von LLMs und anderen modernen Technologien kann die automatische Erzeugung von Inhaltsangaben sowie die Extraktion von Schlagworten, Personen- und Ortsangaben erheblich verbessert werden. Damit lassen sich klassische NLP-Arbeitsschritte wie bspw. die Named Entity Recognition (NER) für Personen oder Ortsnamen vereinfachen (Wettlaufer, 2023, 9-10, 13-14 und Dalfsen et al., 2024).

Praxisbeispiele und Evaluierung neuer Modelle

Anhand von Praxisbeispielen werden wir im Workshop zeigen, wie die neuen convolution-augmented Transformer Modelle in der Praxis funktionieren. Wir werden die Funktionsweise dieser Modelle eingehen und die bisherigen Ergebnisse mit den neuen Modellen vergleichen, um den tatsächlichen Qualitätsgewinn zu bewerten. Darüber hinaus werden wir den Einsatz von modernen LLMs zur Korrektur und Aufwertung der Texterkennungsergebnisse evaluieren. Hierbei werden die Möglichkeiten und Grenzen des Einsatzes von LLMs wie ChatGPT zur Nachbearbeitung und Verfeinerung der Ergebnisse ausgelotet.

Qualitätskontrolle, Aufwertung und Natural Language Processing mit LLM-Systemen

Ein wichtiger Aspekt der Texterkennung ist die Qualitätssicherung der Ergebnisse. Im Workshop wird gezeigt, wie ein Workflow zur Qualitätskontrolle und Aufwertung aussehen kann. Dabei soll die Frage geklärt werden, ob und wie Fehler, die bei konventioneller Texterkennung häufig auftreten, durch den Einsatz von LLM-Systemen minimiert werden können. Ebenso können normalisierte Ausgaben, wie z. B. die Erkennung des langen S (f) als kurzes S (s), mit Hilfe dieser modernen Techniken denormalisiert werden. Der Einsatz entsprechender Systeme könnte somit zu zuverlässigeren, höherwertigen Ergebnissen und einer effizienteren Verarbeitung großer Textmengen führen. Doch gilt auch hier: Vertrauen ist gut, Kontrolle ist besser!? Lässt sich die kreative Natur der LLMs soweit zähmen, dass die Ergebnisse automatisch weiterverarbeitet werden können? Der vollautomatische Einsatz von LLM-basierten Werkzeugen und der „Editor-in-the-loop“-Ansatz werden gegenübergestellt und ihre Vor- und Nachteile beleuchtet. Der "Editor-in-the-loop"-Ansatz bietet eine höhere Sicherheit bei der Qualität der Überarbeitung durch den Einsatz zusätzlicher Ressourcen (Pollin, 2024). Die automatisch er-

zeugten Überarbeitungen werden in einem weiteren manuellen Bearbeitungsschritt kontrolliert bzw. der Abgleich durch den Editor bei der Zusammenführung unterstützt. Darüber hinaus wird die Konkurrenzfähigkeit zu aktuellen NLP-Technologien zur Identifikation und Extraktion von Entitäten und Schlüsselkomponenten demonstriert (Armaselu, 2024, 1-4).

Interaktive Texterschließung

Ein weiterer Schwerpunkt des Workshops sind die neuesten Entwicklungen von LLM-Systemen zur interaktiven Texterschließung. LLM-Systeme können mit ihren eigenen Datensätzen „chatten“, wobei die Qualität stark vom Kontextfenster des jeweiligen Systems abhängt. Dies ermöglicht eine interaktive und explorative Texterschließung, die weit über bisherige Methoden der Informationsverarbeitung hinausgeht. Der Benutzende kann durch gezielte Fragen tiefere Einblicke gewinnen und Muster und Zusammenhänge erkennen, die dem menschlichen Auge normalerweise verborgen bleiben (Wettlaufer, 2023, 5 und 13). Wir werden diskutieren, wie diese Technologien dazu beitragen können, die Qualität und Genauigkeit von Texterkennungssystemen zu verbessern und neue Anwendungsmöglichkeiten zu erschließen, und ob die Ergebnisse den Anforderungen detailorientierter geisteswissenschaftlicher Forschung gerecht werden.

Praktische Anwendung und Nutzen für die Teilnehmenden

Im Workshop lernen die Teilnehmenden die Texterkennungs- und Transkriptionsplattform eScriptorium kennen, mit deren Hilfe wir zunächst die Transkriptionen als Grundlage für die weitere Bearbeitung erstellen. Vom Upload der Dateien über die Layout- und Texterkennung bis hin zur Annotation und dem Datenexport werden die Teilnehmenden mit den erweiterten Funktionen von eScriptorium vertraut gemacht. Anschließend werden wir die erzeugten Daten mit Hilfe von LLMs weiterverarbeiten. LLMs bieten hierbei ein breites Anwendungsfeld für die weitere Anreicherung der Texte wie NER sowie verschiedener Post-Correction-Szenarien. Aus Zeitgründen werden wir uns im Workshop bei den Post-Correction-Szenarien auf die Nachkorrektur von OCR-Fehlern limitieren, um dieses Thema möglichst eingehend behandeln zu können. Hierzu gehören typischerweise falsch erkannte Buchstaben bzw. alte Buchstaben, die in moderne Buchstaben übersetzt wurden, die Ihnen optisch ähnlich sind (Bsp. f und f) und fehlerhaft wiedergegebene Eigennamen.

Der Ablauf des halbtägigen Workshops würde grob folgendermaßen aussehen (Pause nach ca. zwei Stunden):

1 Std. Arbeiten mit eScriptorium

30 Min. Ausblick auf die kommenden Transformer-Modelle und ihre Fähigkeiten

1 Std. Post-Correction der OCR-Ergebnisse mit Hilfe von LLMs

45 Min. Weiterverarbeitung der Inhalte mittels LLMs (NER, Text Analysis (Summarization))

30 Min. Verweis auf rechtliche Fragen und Abschlussdiskussion

Unser Ziel ist es, den Teilnehmenden ein Verständnis für die neuesten technologischen Fortschritte in der Künstlichen Intelligenz zu vermitteln und deren praktischen Nutzen für die Texterkennung und -analyse aufzuzeigen. Die Teilnehmenden sollen in die Lage versetzt werden, diese Technologien in ihren eigenen Projekten anzuwenden und die Vorteile einer verbesserten digitalen Texterkennung voll auszuschöpfen. Hierbei werden wir auch auf die Integration von LLM-Systemen in bestehende Arbeitsabläufe eingehen und praktische Tipps zur Implementierung und Nutzung dieser Technologien geben. Hierbei werden wir neben den etablierten, kommerziellen LLM-Systemen wie ChatGPT auch mindestens einen weiteren kleineren open-source Systeme betrachten. Die im Sinne einer guten wissenschaftlichen Praxis auch eine höhere Nachvollziehbarkeit und Reproduzierbarkeit gewährleisten. Neben dem praktischen Umgang mit LLM-Systemen werden auch rechtliche Aspekte angesprochen, die es zu beachten gilt.

Das GitHub-Repo, mit dem im Workshop verwendeten Materialien (Skripte/Notebooks) werden den Teilnehmenden zur Verfügung gestellt.

Fazit

Die neuesten Entwicklungen im Bereich der Künstlichen Intelligenz und eScriptorium bieten spannende Möglichkeiten für die Texterkennung und -verarbeitung historischer Dokumente. Mit der Einführung von Transformer-Modellen in Kraken und dem Einsatz von LLMs können die Grenzen der bisherigen Texterkennung überwunden und neue Wege der Informationsverarbeitung erschlossen werden. In unserem Workshop werden wir diese Technologien näher beleuchten, ihre Anwendungsmöglichkeiten diskutieren und praktische Beispiele vorstellen. Vorkenntnisse in den verschiedenen Bereichen werden nicht benötigt.

Zu den Organisator*innen des Workshops

Larissa Will war Projektmitarbeiterin im Projekt OCR-BW an der UB Mannheim. Sie ist dort weiterhin im Bereich OCR und Forschungsdatenmanagement tätig als Referentin für Forschungsdatenmanagement und Digitalisierung und berät Forschende aus den Geisteswissenschaften vor allem im Bereich OCR mit Fokus auf eScriptorium.

Jan Kamlah ist seit 2017 Projektmitarbeiter und Softwareentwickler in verschiedenen Projekten (u.a. OCR-D Modulprojekt und BERD@NFDI) mit Schwerpunkt moderne Texterkennung und Datenextraktion an der UB Mannheim.

Thomas Schmidt war Projektkoordinator im DFG-Projekt „OCR-D: Workflow für werkspezifisches Training auf Basis generischer Modelle mit OCR-D sowie Ground-Truth-Aufwertung“ an der UB Mannheim. Mittlerweile ist er im Forschungsdatenzentrum und beschäftigt sich mit Data Literacy.

Dr. Sarah Lang ist Postdoc am „Zentrum für Informationsmodellierung“ der Universität Graz. Zu ihren Forschungsinteressen gehören Distant Reading, Distant Viewing und die Nutzung von Methoden der Computational Humanities für die Geschichte des alchemistischen und chemischen Drucks.

Dorothee Huff war seit 2019 Projektmitarbeiterin im Projekt OCR-BW an der UB Tübingen und weiterhin für den OCR-Service der UB Tübingen mit Schwerpunkt Transkribus zuständig. Zudem arbeitete sie als wissenschaftliche Mitarbeiterin im DFG-Projekt „Narrative Vermittlung religiösen Wissens“. Mittlerweile ist sie Leiterin der Abteilung Handschriften und Historische Drucke mit Restaurierungswerkstatt und Digitalisierungszentrum an der UB Tübingen.

Fußnoten

1. Drop-in-Replacement bezeichnet die Fähigkeit einer Softwarekomponente oder eines Bauteils, eine bestehende Komponente zu ersetzen und mit dieser vollständig kompatibel zu sein.

Bibliographie

Armaselu, Florentina. 2024. “Small-Scale Testing on Generative AI and Post-OCR Correction in Historical Datasets.” In Digital Humanities Benelux 2024 Conference. <https://zenodo.org/records/11403647> (zugegriffen: 23. Juli 2024).

“**ChatGPT.**” n.d. Chatgpt.com. <https://chatgpt.com/> (zugegriffen: 23. Juli 2024).

Dai, Zihang, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le und Ruslan Salakhutdinov. 2019. “Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context.” In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2978–88.

Dalfsen, Arjan van, Folger Karsdorp, Ayoub Bagheri, Thirza van Engelen, Dieuwertje Mentink, und Els Stronks. 2024. „Direct and Indirect Annotation with Generative AI: A Case Study into Finding Animals and Plants in Historical Text“. In Proceedings of the Computational Humanities Research Conference, 2024. Aarhus, Denmark. <https://ceur-ws.org/Vol-3834/paper74.pdf>.

“**Introducing Transkribus super models – get access to ‘The Text Titan I.’**” 2023. READ-COOP. READ-COOP SCE. 10. Juli 2023. <https://readcoop.eu/de/introducing-transkribus->

super-models-get-access-to-the-text-titan-i/ (zugegriffen: 23. Juli 2024).

Kiessling, Benjamin. n.d. Conformer_ocr: Text Recognizer with a Conformer. https://github.com/mittagesen/conformer_ocr (zugegriffen: 23. Juli 2024).

“Kraken — Kraken Documentation.” n.d. Kraken.Re. <https://kraken.re/main/index.html> (zugegriffen: 23. Juli 2024).

Pollin, Christopher. 2024. Workshopreihe "Angewandte Generative KI in den (digitalen) Geisteswissenschaften" (v1.1.0). Zenodo. <https://doi.org/10.5281/zenodo.10647754> (zugegriffen: 23. Juli 2024).

Stokes, Peter A., Benjamin Kiessling, Daniel Stökl Ben Ezra, Robin Tissot und El Hassane Gargem. 2021. “The EScriptorium VRE for Manuscript Cultures.”, herausgegeben von Claire Clivaz und Garrick V. Allen. In *Classics@ Journal, Ancient Manuscripts and Virtual Research Environments* 18 (1). <https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/> (zugegriffen: 23. Juli 2024).

Wettlaufer, Jörg. 2023. “Methoden der Digital History/Digitalen Geschichtswissenschaft.” In *Handbuch Methoden der Geschichtswissenschaft*, 1–19. Wiesbaden: Springer Fachmedien Wiesbaden.

Wikipedia Beitragende. n.d. “Transformer (Maschinelles Lernen).” Wikipedia, The Free Encyclopedia. [https://de.wikipedia.org/w/index.php?title=Transformer_\(Maschinelles_Lernen\)&oldid=247001526](https://de.wikipedia.org/w/index.php?title=Transformer_(Maschinelles_Lernen)&oldid=247001526) (zugegriffen: 23. Juli 2024).

Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, et al. 2023. “A Survey of Large Language Models.” arXiv [Cs.CL]. <http://arxiv.org/abs/2303.18223> (zugegriffen: 23. Juli 2024).