Niedersächsische Akademie der Wissenschaften zu Göttingen

AKADEMIE DER WISSENSCHAFTEN IN HAMBURG

KI-Methoden im Akademienprogramm: Potenziale und Anwendungsszenarien

Themenkonferenz, durchgeführt von der Akademie der Wissenschaften in Hamburg und der Niedersächsischen Akademie der Wissenschaften zu Göttingen
23. bis 25. September 2024

Veranstaltungsort: Universität Hamburg, Ostflügel, HS 221
Edmund-Siemers-Allee 1, 20146 Hamburg

Montag, 23. September 2024

ab 12:00 Ankunft der Teilnehmenden
12:45 Begrüßung

Workshop

13:00 – 17:00 Jan Kamlah, Thomas Schmidt (Mannheim), Raum 121
eScriptorium

18:00 **Öffentlicher Abendvortrag**, Prof. Dr. Chris Biemann (Hamburg), Raum221
Where do you come from, ChatGPT? – Über Entstehung und Zukunft Großer Sprachmodelle

Dienstag, 24. September 2024

9:00 – 10:30 **Integration von KI in Akademievorhaben: Projekte und Perspektiven**

Theresa Baumann, Matteo Burioni, Max Kristen (München)
Vorüberlegungen zu KI-Anwendungen im Corpus der barocken Deckenmalerei
Das Corpus der barocken Deckenmalerei in Deutschland erforscht und publiziert die Decken- und Wandmalerei der Zeit zwischen etwa 1550 und 1800 auf dem Gebiet der Bundesrepublik Deutschland. Der Beitrag gibt Einblicke in geplante KI-Anwendungen auf diesem Feld.

Wolfram Enßlin, Nathanael Philipp & Nadine Quenouille (Leipzig)
Forschungsportal BACH – Möglichkeiten der Integration von KI-gestützten Methoden
Der Vortrag führt zunächst in aller Kürze in das noch junge Projekt „Forschungsportal BACH“ ein. Daraufhin wird das Augenmerk auf die technische Seite gelenkt. Hierbei werden Möglichkeiten erörtert, wie sich KI-gestützte Methoden in einzelne Bereiche des Projekts integrieren lassen.

Jonatan Jalle Steller, Dominik Kasper (Mainz)
KI × DH: Zum Umgang mit ML, Transformern und LLMs in der Digitalen Akademie
Der Beitrag stellt Leitlinien dar, die die Digitale Akademie der Akademie der Wissenschaften und der Literatur Mainz zu ihrem eigenen Umgang mit den im Titel genannten Technologien entwickelt. Dabei geht es sowohl um eine kritische Einordnung der Technologien und der sie umgebenden Diskurse als auch um konkrete Regeln im Umgang damit. In einem dritten Schritt werden konkrete Einsatzszenarien in den von uns betreuten Projekten vorgestellt, die wir als Abteilung gemeinsam ausgearbeitet haben und derzeit mit Leben füllen.

10:30 – 11:00 Kaffeepause

11:00 – 12:30 **OCR, HTR, Editionen und Urkunden**

Daniel Kinitz (Leipzig)
ML-basierte Texterkennung arabographischer Handschriftenkataloge – Herausforderungen und Best Practices
Die automatische Texterkennung arabischer Schrift liefert erst seit der freien Verfügbarkeit von Machine-Learning-basierten Anwendungen zufriedenstellende Ergebnisse. Dennoch gibt es eine Reihe von Herausforderungen, die im produktiven Einsatz zu bewältigen sind. Der Vortrag stellt Beispiele und Best Practices aus dem Projekt „Bibliotheca Arabica“ vor, in dem die Texterkennung von gedruckten Handschriftenkatalogen Teil des Workflows ist.

Frederik Skidzun (Berlin)
Automatisierte Übersetzung von Urkundenregesten mit DeepL
Die Berliner Arbeitsstelle der Regesta Imperii (Regesten Kaiser Friedrichs III.) führt für das Onlineangebot RI Online die automatisierte Übersetzung kurzer Urkundenregesten durch. Der Beitrag soll einige dabei häufig auftretenden Probleme darstellen und erörtern, wie sich Übersetzungsergebnisse von akzeptabler Qualität erreichen lassen.

Jörg Wettlaufer (Göttingen)
Custom-GPTs für die Entitäten-Erkennung und Auszeichnung
Im Zuge der digitalen Bereitstellung von Editionen und Itineraren historischer Reiseberichte hat sich die Aufgabe ergeben, Reiserouten in hunderten von Texten zu erkennen und auszuzeichnen, damit diese z.B. über digitale Karten visualisiert werden können. Obwohl die zuverlässige Erkennung von Named Entities (NER) ein wichtiger Vorverarbeitungsschritt ist, reicht die einfache Auszeichnung von Orts- und Personennamen allein für die Erkennung von Itineraren nicht aus, da in den Texten häufig auch Orte genannt werden, die nicht auf der Reiseroute liegen. Zu diesem Zweck wird ein Ansatz präsentiert, in dem sowohl die Erkennung von Orten und Personen als auch die Beziehungen zwischen Personen, Orten und Reisebeteiligten Beachtung finden. Hierzu wurden zunächst Annotationsrichtlinien zur Auszeichnung der oben genannten Beziehungen unter Berücksichtigung der örtlichen und zeitlichen Beziehung einer Person zu einem gegebenen Ort erstellt. Anschließend wurde getestet, in welchem Umfang die Annotation dieser komplexen Aufgabe durch Large Language Models unterstützt werden kann. Zunächst wurden die Richtlinien an den Data Analyzer von Chat-GPT 4.0 übergeben und um die Auszeichnung eines Beispieltextes gebeten. In einem weiteren Schritt wurde ein OpenAI Custom-GPT nicht nur mit den Annotationsrichtlinien ausgestattet, sondern auch mit manuell annotierten Beispieltexten trainiert. Schließlich wurde das GPT um die Fähigkeit erweitert, den Geocode der annotierten Orte des Itinerars einzufügen und die Reiseroute auf einer Karte (außerhalb des Chats) anzeigen zu lassen. Der Beitrag präsentiert und diskutiert die Fähigkeit des trainierten GPTs zur automatisierten Auszeichnung/Visualisierung der Personen, Orte und Relationen nach den vorgegebenen Annotationsrichtlinien für Reiseberichte.

12:30 – 14:00 Mittagspause

14:00 – 15:30 **Lexika, Übersetzungen und Annotation**

Jan Christian Schaffert (Göttingen)
Edition, KI und Lexikographie – Wie können digitale und interdisziplinäre Zugänge zu der frühneuhochdeutschen Textwelt geschaffen werden?
Gegenstand des Beitrags ist die Entwicklung und das Training eines Large Language Model (LLM), das das Frühneuhochdeutsche zunächst lemmatisieren sowie bestmöglich semantisieren kann, mit dem Ziel, diese Ergebnisse in digitale Editionen einzuspeisen. Den Rahmen hierfür bilden aktuelle Arbeiten am Frühneuhochdeutschen Wörterbuch (FWB), einem historischen Sprachstadienwörterbuch, das den Wortschatz des Frühneuhochdeutschen (1350-1650) umfassend erfasst und semantisch beschreibt. Die Aufgabe erweist sich aufgrund der fehlenden Normierung und Orthographie des Frühneuhochdeutschen als hochgradig komplex. Erste Machbarkeitsstudien dienen als Grundlage, um im weiteren Vorgehen ein LLM zu entwickeln, das nicht nur für das Frühneuhochdeutsche, sondern auch für andere Low Resource Sprachen relevant sein kann.

Manuel Raaf (München)
Die Lemmatisierung von Zettelkästen mittels Deep Learning: Handschriftenerkennung im Fränkischen Wörterbuch
Für die Lemmatisierung der digitalisierten Zettelkästen des „Fränkischen Wörterbuchs“ wurden zunächst die Reiterkarten aus dem Material von über 790.000 Digitalisaten gefiltert, um anschließend deren handschriftlich eingetragene Lemmata zu erkennen. Hierfür wurden jeweils Methoden des Deep Learnings eingesetzt. Der Vortrag skizziert beide Vorgänge, präsentiert die Ergebnisse und diskutiert diese kritisch. Der Vorgang der Lemmatisierung wird zudem verglichen mit dessen manueller Durchführung durch Hilfskräfte.

Patrick D. Brookshire (Mainz)
Namen erkennen und klassifizieren. Fine-Tuning von Transformer-Modellen
In vielen Projekten des Akademienprogramms werden Textpassagen manuell annotiert oder ganze Datensätze kategorisiert, was beides typische Anwendungsfälle für Transformermodelle sind, die seit 2017 wegen ihrer Performanz nicht nur den NLP-Bereich dominiert haben. Diese Modelle sind die Basis der bekannten generativen LLMs, aber durch eine i.d.R. geringere Modellgröße sowie den Fine-Tuning-Ansatz weniger ressourcenintensiv und leichter zu evaluieren. Darüber hinaus eignen sie sich besonders für eine automatische Vorauszeichnung auf Basis bestehender Projektdaten, wie Beispiele zur Erkennung und Klassifikation von (Personen-)Namen aus Projekten der Akademie der Wissenschaften und der Literatur | Mainz zeigen.

15:30 – 16:00 Kaffeepause

16:00 – 17:30 **Retrieval Augmented Generation und Linked Open Data**

Bärbel Kröger, Bashar Jaan Kahn (Göttingen)
Informationsextraktion aus lateinischen Texten des Repertorium Germanicum mittels Custom GPTs
Die Germania Sacra arbeitet im Bereich der mittelalterlichen Kirchengeschichte, einem Wissensgebiet, in dem die Anwendung von LLMs bisher noch unterrepräsentiert ist. Der Beitrag stellt experimentelle Ansätze vor, die darauf abzielen, ob und inwiefern mit der Anreicherung eines Modells durch die umfangreichen Datenkorpora und Veröffentlichungen der Germania Sacra brauchbare Ergebnisse erzielt werden können.
So wird beispielsweise geprüft, inwiefern es möglich ist, Informationen aus der auf lateinischem Text (sogenannten Regesten) beruhenden Datenbank "Repertorium Germanicum" zu extrahieren, indem das Modell zusätzlich mit Informationen aus strukturierten Projektdatenbanken der Germania Sacra sowie den Citizen Science Projekten FactGrid bzw. Wikidata angereichert wird, die dasselbe Wissensgebiet abdecken.

Thomas Eckart, Felix Helfer, Uwe Kretschmer (Leipzig)
Machine-learning gestütztes Entity Linking
Entity Linking dient der Verknüpfung von Vorkommen benannter Entitäten (wie Personen, Orte, Organisationen) in Texten mit ihren entsprechenden Einträgen in Normdateien. Es werden Experimente zu verschiedenen Ansätzen vorgestellt, die sowohl klassische Embedding-basierte Verfahren, als auch Transformer-basierte Sprachmodelle nutzen. Dabei werden Herausforderungen und Chancen der Aufgabe diskutiert.

Timm Lehmberg, Stefano Valente (Hamburg)
Retrieval Augmented Generation: Anwendungsbeispiel für dokumentbasiertes Parsing
Dieser Beitrag untersucht das Potenzial von Retrieval-Augmented Generation (RAG) für die Datenanalysen im Bereich der Digital Humanities. RAG kombiniert die Fähigkeiten von Retrieval-Systemen mit den generativen Fähigkeiten großer Sprachmodelle, sodass präzisere und kontextbezogene Antworten auf komplexe Anfragen geliefert werden klönnen. Im Rahmen des Hamburger Langzeitprojekts 'Etymologika', wurde ein Use Case entwickelt, bei dem strukturierte textuelle Daten aufbereitet und in ein RAG-System integriert wurden. Der Beitrag beleuchtet die methodischen Schritte zur Vorbereitung und Integration der Daten sowie die dadurch entstandenen neuen Zugänge zu den Daten des Projektes.

19:30 Angebot einer Lichterfahrt auf der Elbe (Selbstzahler)

<u>Mittwoch, 25. September 2024</u>

Hands-On-Sessions

9:00 – 13:00 Timm Lehmberg (Hamburg), Raum 121
Hackathon “Chat mit RAGgate”. Eigene Datenbasen mithilfe von Retrieval Augmented Generation zugänglich machen

9:00 – 13:00 Ines Röhrer (München), Raum 222
Prompt-a-thon “Mein Bot versteht mich nicht!” Verarbeitung von Prompts verstehen und bessere Ergebnisse erzielen