Large Language Models, oder weshalb wir künstliche Intelligenz im Archiv finden sollten

Von Tobias Hodel[1]

Künstliche Intelligenz oder spezifischer sogenannte maschinelle Lernverfahren, die mit Textdaten umgehen können, sind keine Erfindung der 2020er Jahre. Bereits seit Jahren werden sogenannte Sprachmodelle eingesetzt, um Texte aufzubereiten oder Suchen zu verbessern. Mit Anwendungen wie ChatGPT und Suchsystemen, die auf direkter Interaktion basieren, erfahren aber eine Vielzahl von Nutzenden Formen des maschinellen Lernens. Damit wird diese Form der künstlichen Intelligenz ins Bewusstsein vieler Anwenderinnen und Anwender gebracht. Ebenso wird für die Recherche und für die Analyse eine Erwartungshaltung bezüglich der Interaktionsformen mit Textdaten geweckt.

Die Rolle von Archiven in der Anwendung generativer künstlicher Intelligenz ist noch mehrheitlich undefiniert, obwohl bereits eine Vielzahl von Archiven Erfahrungen mit maschinellem Lernen machen, etwa in Form von Anwendungen im Bereich der Texterkennung oder der Suche in Bildbeständen.[2] Wie sollen sich Archive positionieren und mit künstlicher Intelligenz interagieren? Das ist die Ausgangsfrage der folgenden Seiten.

Anwendungen wie ChatGPT existieren erst seit 2022, weshalb es vermessen wäre, bereits empirisch gesättigte Erkenntnisse zu versprechen. Bereits kurze Zeit nach der Einführung der Chatsysteme kann jedoch aufgezeigt werden, wie die neu verfügbaren Werkzeuge auch oder vielmehr insbesondere in Archiven eingesetzt werden könnten. Der vorliegende Beitrag versucht aus diesem Grund, einen Aufriss der Thematik mit Blick auf Herausforderungen in Archiven zu leisten.

Der Startpunkt ist nicht die Technologie, sondern das Problemgemenge, das wir in Archiven identifizieren. Dabei geht es nicht darum, ein Defizit zu konstatieren, sondern vielmehr aufzuzeigen, weshalb Daten aus Archiven besondere Herausforderungen an Technik und Mensch stellen. Davon ausgehend kann mit Blick auf die Technologie aufgezeigt werden, wie Archive ein Korrektiv darstellen könnten, das auch auf die entstehenden Modelle zurückwirkt. So kommen wir schließlich zu einem Ende, das mögliche Einsatzszenarien für große Sprachmodelle skizziert und eine Erprobung anregen soll.

---

[1] Die Ausführungen basieren auf mehr oder minder systematischen Versuchen mit Large Language Models und Erfahrungen im Umgang mit Sprachmodellen. Der Autor möchte die frühe Phase der Auseinandersetzung mit LLMs betonen. Aktuell können wir noch nicht abschätzen, inwiefern sich Risiken und Chancen die Balance halten. Alle Links wurden letztmals am 31. 10. 2023 abgerufen.

[2] Siehe auch den Beitrag von Florian *Spiess* in diesem Band.

Bevor wir uns den spezifischen Herausforderungen widmen, sollen zunächst einige Gedanken zur sprachlichen Beschreibung der Technologie formuliert werden. Der Begriff der künstlichen Intelligenz ist noch immer unterdefiniert. Der Terminus ist keine *per se* besonders neue Erfindung, sondern wird schon seit mehr als 60 Jahren als eigenes Forschungsfeld bearbeitet. Mit *künstlicher Intelligenz* ist denn auch nicht eine technische Herangehensweise oder ein Methodenapparat gemeint, sondern vielmehr eine Forschungsrichtung.[3]

Aufgrund der begrifflichen Unschärfe bevorzugen methodisch orientierte Forschende den Fachbegriff des maschinellen Lernens, der Methoden umfasst, die große Datenmengen nutzen, um Verfahren zu entwickeln, die auf Mustern basieren oder die versuchen über Muster Ähnlichkeiten zu eruieren, die nachgeahmt werden können.[4] Maschinelles Lernen ist denn auch der Methodenapparat, der in der Text- und Bildanalyse mit Erfolg seit einigen Jahren eingesetzt wird und vermehrt mit dem Begriff der *künstlichen Intelligenz* verschmilzt.

Aus technischer Warte ist das sogenannte *deep learning* die erfolgreichste Form des maschinellen Lernens, vorausgesetzt, es existiert genügend Material, auf das sich das System berufen kann.[5] Diese Technologien sind allerdings keine neutralen Agenten. Weder die Daten[6] noch die Algorithmen[7] stehen außerhalb unserer Wissenssysteme und -repräsentationen, sondern gehören zu Systematiken, wie wir die Welt verstehen und ordnen. Für die Archivwissenschaften sind das keine bahnbrechenden Neuigkeiten. Die kritische Archivforschung hat in den vergangenen zwei Jahr-

---

3 Siehe leitend Stuart J. *Russell* u. a.: Artificial intelligence: a modern approach. Harlow 2022 (Pearson series in artificial intelligence). S. viii.

4 Siehe einführend und mit Blick auf die Geschichtswissenschaft Tobias *Hodel*: Die Maschine und die Geschichtswissenschaft: Der Einfluss von deep learning auf eine Disziplin. In: Digital History: Konzepte. Methoden und Kritiken Digitaler Geschichtswissenschaft. Bd. 6. Hg. von Karolina Dominika *Döring* u. a. (Studies in Digital History and Hermeneutics) Berlin, Boston 2022. S. 65–80. https://doi.org/doi:10.1515/9783110757101-004.

5 Es wird von sogenannten Trainings-, Validierungs- und Testsetdokumenten ausgegangen. Siehe anhand des Beispiels der Texterkennung, in der dieser Prozess ebenfalls genutzt wird. Siehe mit Bezug zur Texterkennung Tobias *Hodel*: Konsequenzen der Handschriftenerkennung und des maschinellen Lernens für die Geschichtswissenschaft. Anwendung, Einordnung und Methodenkritik. In: Historische Zeitschrift 316/1 (2023) S. 151–180. https://doi.org/10.1515/hzhz-2023-0006.

6 Siehe besonders instruktiv: *Raw Data* Is an Oxymoron. Hg. von Lisa *Gitelman* (Infrastructures series). Cambridge 2013.

7 Siehe zu Beeinflussungen anhand von Sexismus Mar *Hicks*: Sexism Is a Feature, Not a Bug. In: Your computer is on fire. Hg. von Thomas S. *Mullaney* u. a. Cambridge, Massachusetts 2021. S. 135–158. https://doi.org/10.7551/mitpress/10993.003.0011. – Allgemein einführend siehe Safiya Umoja *Noble*: Algorithms of Oppression: How Search Engines Reinforce Racism. New York 2018.

zehnten wiederholt aufgezeigt, inwiefern die Art und Weise des Umgangs mit Archiven Vorstellungen transponiert und unser Umgang mit Archivalien die Deutung derselben beeinflusst.[8]

Verkürzt: Funktionieren der künstlichen Intelligenz

Die *deep learning* Systeme, mit denen wir mehr und mehr täglich umgehen, basieren in erster Linie auf großen Mengen an strukturierten oder unstrukturierten Daten. Um beispielsweise erfolgreich Texterkennung zu betreiben, ist es nötig, dass umfangreiche Beispiele (im Sinne von Trainingsdaten) für die Algorithmen zur Verfügung gestellt werden.[9] Für Sprachmodelle, wie sie in ChatGPT verbaut sind, sind gleichfalls umfangreiche Textvorlagen notwendig, bei denen die gesamte Wikipedia einen winzigen Ausgangspunkt bildet, um Modelle zu erstellen.[10]

Die dabei erzeugten Modelle sind nicht viel anderes als optimierte Repräsentationen des visuellen Inputs (Texterkennung) oder des Verhältnisses der Daten zueinander (Sprachmodelle). Sprich im *deep learning* wird versucht, einen optimalen Status zu erzeugen, um beim nächsten Schritt ähnliche Resultate zu generieren. In diesen so generierten Modellen ist immer nur so viel *Weltwissen* vorhanden, wie in den Trainingsmaterialien vermittelt wurde. Wobei die Optimierung immer nur auf ein Ziel hin läuft, das aus der Erkennung oder der Generierung von Text besteht, analog zu den Materialien im Training. Sinn oder gar Kreativität sollte darin nicht gesucht werden.

Im Verlauf der vergangenen Jahre waren die Entwicklerinnen und Entwickler der Sprachmodelle selbst von den Resultaten überrascht, die die imitierenden Systeme zur Generierung von Text hervorbrachten. Ursprünglich waren die Modelle nicht zur Ausgabe von Fakten oder ähnlichem gedacht, sondern einzig zur Generierung von Textbausteinen, etwa zur Korrektur der Grammatik in existierenden Texten. Als Reaktion auf die gefühlt gehaltvollen und intelligenten Texte wurde den Modellen teilweise eine Reflexionsfähigkeit und gewissermaßen eine Intelligenz

---

8 Siehe beispielsweise Ann Laura *Stoler*: Along the archival grain: Epistemic anxieties and colonial common sense. Princeton, NJ Oxford 2010 und Eric *Ketelaar*: Tacit narratives: The meanings of archives. In: Archival Science 1/2 (2001), S. 131–141. https://doi.org/10.1007/BF02435644.

9 Siehe beispielsweise in diesem Band den Beitrag von Dorothee *Huff* oder wie dargelegt in Guenter *Muehlberger* u. a.: Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. In: Journal of Documentation 75/5 (09.09.2019) S. 954–976. https://doi.org/10.1108/JD-07-2018-0114.

10 Der Ausgangspunkt ist in vielen Fällen ein sogenannter *scrape* von Texten aus dem Internet. Zusätzlich kann auch gerade in den kommerziellen Systemen nachgewiesen werden, wie urheberrechtlich geschützte Daten in die Modelle eingeflossen sein müssen. Siehe dazu Kent K. *Chang* u. a.: Speak, Memory: An archaeology of books known to ChatGPT/GPT-4, 28.04.2023. https://doi.org/10.48550/arXiv.2305.00118.

zugeschrieben, die den Blick auf die eigentlichen Fähigkeiten der Systeme jedoch vollends verstellt.[11]

In der Anlage handelt es sich auch bei den großen Sprachmodellen, den sogenannten *Large Language Models*, um Systeme, die auf die Generierung von Text trainiert wurden und in einem zweiten (aufwändigen) Schritt lernten, welche *sinnvollen* Antworten zu Anfragen in ihrem Modell gefunden werden können. Problematisch wird dieser Umstand, sobald das Modell mit Fragen konfrontiert wird, die es nicht oder nur selten in seinen Trainingsmaterialien gesehen hat. Aufgrund des gelernten Umgangs mit ähnlichen Fragen werden sodann Antworten erzeugt, die sich zwar glaubhaft anhören, jedoch eine inhaltlich erfundene Zusammensetzung mit plausibler Form sind.

Anhand eines Vorzeigebeispiels lässt sich dieses Problem aufzeigen, in dem eine Anfrage gestellt wird, welche Preise eine Person gewonnen hat, die in einem Feld (beispielsweise einer Disziplin) aktiv ist. Aufgrund der Frage *Welche Preise hat X gewonnen* versucht das Modell eine Antwort zu generieren, wie sie im Trainingsmaterial vorgekommen ist. Dabei wird auch der zusätzliche Input (Name von X sowie beruflicher Hintergrund) miteinbezogen. Vielfach spuckt das *Large Language Model* dann eine Vielzahl an echten und erfundenen Preisen aus, unabhängig davon, ob die Person den Preis gewonnen hat.

Parallel dazu sehen wir alle Voreingenommenheit widergespiegelt, die das Modell in den Trainingseinheiten *gesehen* hat. Explizit Minderheiten und sozial schwächere Gruppen sind in diesen Trainingsmaterialien oftmals unter- oder falsch repräsentiert, und dementsprechend beeinflusst fallen die Modellresultate aus. Die Ausgaben können jederzeit in rassistische, sexistische oder anderweitig problematische Bereiche kippen.[12]

Im Wissen um diese Einschränkung lässt sich relativ schnell ein guter Umgang mit den Algorithmen finden. Als unabhängige, Wissen wiedergebende Textgeneratoren eignen sie sich folglich nicht. Andererseits werden die Systeme enorm wirkmächtig, sobald es darum geht, Informationen aus Textdaten zu extrahieren (wer wird in einem Text erwähnt, welche Themen werden angesprochen), oder gar Inhalte zusammenzufassen. Das gilt natürlich auch oder insbesondere für Archive.

Eine Komponente und gleichzeitig eine neue Kompetenz wird das sogenannte *prompt engineering*. Darunter versteht man möglichst hilfreiche Anfragen, die an die Chatsysteme oder allgemein die Sprachmodelle gestellt werden. Dabei lohnt sich eine sorgsame Texteingabe, um die Anfragen

---

[11] Siehe dazu den Fall eines Entwicklers, der dem LaMDA System/Sprachmodell Bewusstsein attestierte: Nico *Grant*, Cade *Metz*: Google sidelines engineer who Claims its A.I. is sentient. In: The New York Times (12. 06. 2022). https://www.nytimes.com/2022/06/12/technology/google-chatbot-ai-blake-lemoine.html.

[12] Siehe leitend: Emily M. *Bender* u. a.: On the dangers of stochastic parrots: can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). New York 2021. S. 610–623. https://doi.org/10.1145/3442188.3445922.

zu verbessern. Zentral bei den Anfragen ist die präzise Definition der Aufgabe (*Zusammenfassung verfassen*, *Stichwörter identifizieren*, *Text orthographisch korrigieren*) und die Benennung einer Perspektive (*für Grundschüler*, *für ein akademisches Publikum*). Hilfreich sind auch Beispiele, die in einer Anfrage mitgegeben werden. Die Beispiele können etwa aus Listen bestehen, aus denen die Antwort bestehen kann. Um die Rückmeldung noch elaborierter zu gestalten, können auch Definitionen mitgegeben werden, beispielsweise um dem Sprachmodell Datenbankstrukturen zu vermitteln.

Die Herausforderung: Weshalb das Archiv nicht für AI gemacht ist

Archive sind Horte besonders komplexer Daten. Wir finden auf engem Raum unterschiedliche Sprachen und – zumeist wichtiger – unterschiedlich historische Sprachformen. Semantischer Wandel und eine (für moderne Sprachformen ungebräuchliche) komplexe Syntax sind dabei nur Teile der Herausforderung. Die Historizität der Dokumente und die je nach Archiv stark divergierenden Fachsprachen sind ein weiterer Teil der Herausforderung, die wir in der unendlichen Anzahl von Dokumenten finden: Von der Verwaltungssprache über die Wirtschaftssprache bis hin zu fachwissenschaftlichen Sprachen etwa in Musikarchiven finden wir unzählige spezialisierte Ausprägungen.

Zu den formellen Herausforderungen kommt ein zentrales weiteres Problem: Sprachmodelle wie ChatGPT werden gerne als Wissensspeicher (miss-)verstanden. Gerade für Wissen und Informationen, die im frei zugänglichen Internet nicht hochgradig redundant repräsentiert werden, eignet sich der Einsatz der LLM nicht für die Wiedergabe. Die bereits problematisierten Stereotypisierungen und Vereinheitlichungen sind denn auch in diesem Bereich zu identifizieren. Städtische oder lokale Akteure, die zurecht in den jeweiligen Archiven ein gewisses Renommee genießen, sind in den Sprachmodellen höchstens marginal präsent und jegliche Informationen dazu sehr kritisch zu prüfen.

Es spricht auf den ersten Blick daher nicht viel für den Einsatz von *Large Language Models* in Archiven. Da die genannten Probleme und Herausforderungen von Archivmaterial aber nicht nur für die Algorithmen gelten, sondern auch für Menschen, wird das Auffinden und Einordnen von Archivalien zu einer herausfordernden Aufgabe, derer wir uns auch unter Einsatz von LLMs stellen sollten. Damit ist nicht insinuiert, dass dies die einzige Zugangsform ist, sondern eine von diversen Möglichkeiten, die wir dennoch in Betracht ziehen und kritisch wagen sollten.

Die Chance des Zugangs: Disparates Archivmaterial einordnen

Trotz der genannten Hürden vertrete ich hier die Meinung, wir können und sollen große Sprachmodelle im Archiv nutzen. Dabei sehe ich unterschiedliche Einsatzszenarien, die in aller Kürze skizziert werden sollen. Für Suchvorgänge, die über disparates Material und historische Sprachstufen laufen, eignet sich der Einsatz großer Modelle ebenso, wie für die Erschließung, für das

Vergleichen von Dokumenten oder die Extraktion von Informationen. Im Laufe der umfangreichen Trainingsprozesse haben Sprachmodelle eine Vielzahl von Sprachstufen gesehen und häufig auch historische Sprachen sowie vielfältige Formen der Ähnlichkeit. Diese Ausgangslage sollten wir uns zu Nutze machen.

Eine Volltextsuche in Archivmaterial wird in vielen Fällen keine ausreichenden oder gar befriedigenden Resultate liefern. Sprachmodelle können für diesen Schritt zumindest teilweise Abhilfe schaffen. Wie sich an ersten Beispielen aufzeigen lässt, können die großen Modelle relativ gut mit historischen Sprachstufen und auch mit Mehrsprachigkeit umgehen und vermögen Ähnlichkeiten im Sprachgebrauch dahingehend einzuordnen, dass sie sinnvolle Analogieschlüsse ziehen können. Sowohl nicht mehr gebräuchliche Wortformen als auch Fehllesungen der automatisierten Texterkennung können auf diese Weise eingeordnet werden.

Mit der Unterstützung der Suche einher geht der Support der Erschließungsarbeit, die in vielen Archiven ressourcenintensiv betrieben wird. Gerade für die Tiefenerschließung lassen sich Sprachmodelle einsetzen, wenn die gewünschte Form der Erschließung im *prompt* beschrieben wird. Diese Form des *one-shot learning*, also der Instruktion und Anpassung über die Eingabe, ermöglicht, dass auch der Archivspezifik und eigenen Formen der Erschließung Rechnung getragen wird. Voraussetzung ist die Existenz digitaler Textelemente, wobei neuere Iterationen der LLMs auch bis zu einem gewissen Grad mit Bilddaten umgehen können. Die Extraktion von Schlagworten, genannten Akteuren oder Zusammenfassungen etwa in Form der Regestierung, kann somit automatisiert oder unterstützt durch ein menschliches Korrektiv in Kooperation aufgearbeitet werden.

Ohne die Textgenerierungsfunktionalität von Sprachmodellen zu nutzen, eignen sich die Modelle gut, um Dokumentenvergleiche anzustellen. Einer in Volltext vorliegenden Archivalie kann dadurch ein *nächstähnliches* Dokument zur Seite gestellt werden. Solche Ähnlichkeiten lassen sich gleichzeitig als Visualisierungen ausgeben, womit neue Formen des Zugriffs entstehen.

Aktuell ist es vorwiegend die Kreativität, die bei solchen Informationsextraktionsvorgängen einschränkend wirkt. Dokumente können für die Suchenden übersetzt zur Verfügung gestellt werden. Ein Vorgang, der sich auf immer mehr Webseiten beobachten lässt und idealerweise gekennzeichnet wird. Aus den Übersetzungen oder der ursprünglichen Form können zusammenfassende Einheiten kreiert oder Stichworte erzeugt werden. Wer schon selbst Regesten für Urkunden angefertigt hat, weiß um die Herausforderungen solcher Unterfangen. Mit Sprachmodellen wird die Qualität sicherlich nicht an diejenige von Expertinnen und Experten heranreichen, aber die quantitative Verarbeitung kann auch für Dokumentenformen wie Protokolle oder andere Massenquellen angewandt werden, deren Be- und Verarbeitung nie in Erwägung gezogen werden würden. Um die Technologie zu nutzen, ist es wichtig, über Qualitätsstandards nachzudenken: Der Verlust an (scheinbarer) Perfektion in Erschließung und Verfügbarmachung wird durch die neuen quantitativen Dimensionen zwar nicht wettgemacht, jedoch mindestens ergänzt bzw. erweitert.

In all diese Vorgänge werden sich sowohl kleine Fehler oder Hyper-Normalisierungen aber auch, und das ist gravierender, Beeinflussungen einschleichen. Der schon erwähnte *bias* ist denn

auch die größte Herausforderung im Umgang mit den Systemen.[13] Die Markierung des Einsatzes von großen Sprachmodellen bildet eine zentrale Aufgabe, um zumindest ansatzweise auf die damit einhergehenden Probleme aufmerksam zu machen. Sinnvoll sind auch Anleitungen und Beispiele, die aufzeigen, inwiefern Resultate fehlerhaft und Schlüsse problematisch sein können.

Die Chance der Archive: Ein Korrektiv in der Masse

Ein Archiv ist nicht nur ein Gefäß, um Verwaltung und Politik nachvollziehbar zu machen, sondern kann auch im digitalen Zeitalter eine entscheidende Rolle einnehmen, da die von ihm verwaltete Datenfülle potenziell massive und gleichzeitig gewinnbringende Erweiterungen für die Algorithmen ermöglicht. Das Archiv als Speicher einer Diversität bekommt damit eine weitere Aufgabe und kann gleichzeitig selbst seiner Quellenmassen bis zu einem gewissen Grad Herr werden.

Die offene Publikation der Daten wird gleichzeitig eine geschätzte Ressource, um die aktuell meist geschlossenen Systeme (insbesondere die GPT/ChatGPT Familie von OpenAI) durch offene Entwicklungen zu ergänzen und zu erweitern. Dabei geht es nicht darum, die Monetarisierung zu verhindern, sondern vielmehr, die verwendeten Daten offenzulegen und nachvollziehbar zu machen.

Für die kritische und umsichtige Nutzung der Technologien brauchen wir eine intensive Beschäftigung mit ihren Chancen und Risiken. Nur mit einer solchen *data* und *algorithmic literacy* werden die Vorteile implementiert und Risiken vermindert. Eine solche Sicht kann nicht von Dienstleistungsbetrieben eingebracht werden, sondern kommt optimalerweise aus den Reihen der Archivarinnen und Archivare. Ansonsten fehlt sowohl ein Verständnis für die Hindernisse, die sich im Material finden, als auch ein Abwägen der Informationsfülle, die gefunden werden kann. Mit den neuen Ansätzen werden somit nicht Personen ersetzt, sondern die Beschäftigung mit dem Material verschoben – weg von repetitiven Arbeiten und hin zu Reflexionsakten, die ein vertieftes Verständnis der Materialien verlangen.

Die Textmassen, die sich in den Archivmagazinen anhäufen, bedürfen nicht nur elaborierter Suchen, sondern auch der Kontextualisierung im Wortsinn. Die Textproduktion ist nur in ihrem Kontext zu ergründen und legt Vergleiche und die Bearbeitung über Texte nahe. Dabei sollten jedoch nicht nur Chat-Systeme im Fokus stehen, sondern unterschiedliche Frage- und Auskunftssysteme, die zielgerichtet sind oder aber auf Ähnlichkeiten aller Art abheben.

Letztlich sollten wir aber dennoch nicht den Vorteil beim Umgang mit den Dokumenten mit großen Sprachmodellen vergessen. LLMs eröffnen den Archiven eine Zugriffsform, die bis vor

---

[13] Ein Ansatz sind Modelcards, wie sie mittlerweile auch auf der Plattform Huggingface eingesetzt und in folgendem Artikel ausführlich beschrieben werden: Margaret *Mitchell* u. a.: Model Cards for Model Reporting. In: Proceedings of the conference on fairness, accountability and transparency (29. 01. 2019) S. 220–229. https://doi.org/10.1145/3287560.3287596.

wenigen Jahren undenkbar war und nun immer selbstverständlicher in unterschiedlichen Sparten eingesetzt wird. Archive können und sollen davon profitieren.

Ein überbordender Technologieglaube passt sicher nicht ins Archiv, jedoch sollten neue Technologien auf jeden Fall als Chance verstanden werden, um die Massen an Archivalien neu und anders deuten zu lassen. Deshalb sollten wir nicht vergessen, die Dokumente aus den Archiven in die neuen Systeme einzuspeisen und damit deren Leistungsfähigkeit zu erhöhen.