Kasprzik, Anna

Article — Published Version

Künstliche Intelligenz für die Inhaltserschließung – ein Statusupdate

BuB - Forum Bibliothek und Information

Suggested Citation: Kasprzik, Anna (2024) : Künstliche Intelligenz für die Inhaltserschließung – ein Statusupdate, BuB - Forum Bibliothek und Information, ISSN 1869-1137, Berufsverband Information Bibliothek (BIB), Reutlingen, Vol. 76, Iss. 08-09, pp. 442-445

This Version is available at:
http://hdl.handle.net/11108/627

Anna Kasprzik

Künstliche Intelligenz für die Inhaltserschließung – ein Statusupdate

Wie die ZBW maschinelles Lernen für die automatisierte Inhaltserschließung einsetzt und welche Herausforderungen und Chancen sich daraus ergeben

In den letzten Jahren haben Anwendungen auf der Basis der neusten Generation von Machine-Learning-Modellen wie etwa ChatGPT für einen Hype gesorgt, der alle anderen Hochphasen der Künstlichen Intelligenz (KI) übertroffen hat – insbesondere in der Öffentlichkeit und den Medien. Auch in Bibliotheken nimmt die Automatisierung mit KI-Methoden eine zunehmend bedeutende Rolle ein, und das nicht erst seit ChatGPT. Trotzdem wirft die jüngste Innovationswelle Fragen auf, die in dieser Schärfe bisher nicht gestellt wurden – sind die Aktivitäten von Bibliotheken im Bereich der Erschließung und Metadatenerstellung nun obsolet? Oder stellt dies eine ungeahnte Chance dar, die Aufgaben einer Bibliothek auf ihren Kern zurück- und damit in ein neues Zeitalter zu überführen? Schockstarre oder Reanimation?

1. Die Automatisierung der Inhaltserschließung an der ZBW bis heute

Die inhaltliche Erschließung, also das Annotieren von Publikationen mit semantischen Angaben zur Erhöhung der Auffindbarkeit von relevanter Literatur im Bestand, ist eine der Kernaufgaben der ZBW. Aufgrund der digitalen Publikationsflut ist ein flächendeckendes intellektuelles Annotieren unmöglich geworden, sodass Automatisierung als Strategie naheliegt. Wie bereits in der BuB-Ausgabe 06/2022 (»Schwerpunkt Künstliche Intelligenz«)[1] berichtet, hatte sich die ZBW schon über ein Jahrzehnt mit den Möglichkeiten einer automatisierten Inhaltserschließung befasst, bevor die Entscheidung fiel, die notwendige angewandte Forschung im Machine-Learning-Bereich und die Prototypentwicklung auf Open-Source-Basis im eigenen Haus durchzuführen. 2020 wurde eine Pilotphase für die Überführung der so entstandenen prototypischen Lösungen in einen produktiven Dienst und für den Aufbau einer geeigneten Architektur zur Integration dieses Dienstes in die bestehende Metadateninfrastruktur ausgerufen. Eine erste Version des AutoSE-Dienstes ging 2021 live und mittlerweile ist die Hälfte der englischsprachigen Ressourcen im Bestand (das entspricht einem Viertel des Gesamtbestandes) mit automatisiert erzeugten Verschlagwortungen versehen.[2] Zusätzlich wird der Output über eine Schnittstelle zur Unterstützung der intellektuellen Sacherschließung bereitgestellt, die an der ZBW über die Plattform Digitaler Assistent (DA-3)[3] erfolgt. Der Dienst bedient also sowohl das Ziel einer vollautomatisierten als auch das einer maschinenunterstützten Erschließung – so kann eine Vielzahl von Publikationen verarbeitet werden, die sonst gar nicht inhaltlich erschlossen würden, und die Verschlagwortung von Publikationen, die an der ZBW im Fokus der intellektuellen Inhaltserschließung stehen, geht leichter von der Hand.

Eine Kernkomponente des Dienstes ist das von der Finnischen Nationalbibliothek entwickelte Open-Source-Toolkit Annif, welches verschiedene Standard-Algorithmen für eine automatisierte Inhaltserschließung anbietet, aber auch das Integrieren eigener Machine-Learning-Modelle erlaubt. In AutoSE dient Annif sozusagen als Steckrahmen für eine Kombination mehrerer Modelle – einschließlich einer ZBW-Eigenentwicklung – und wird flankiert von einer Reihe von Mechanismen für wissenschaftliches Experimentieren, Parameteroptimierung, Qualitätskontrolle und den Anschluss an andere Metadatenworkflows im Haus. Das AutoSE-Team beteiligt sich weiterhin an der Weiterentwicklung von Annif, prüft regelmäßig, ob sich Erkenntnisse aus dem AutoSE-Kontext als neue Funktionalitäten in Annif integrieren lassen,[4] berät andere Institutionen zu dessen Funktionsweise und veranstaltet zusammen mit der Finnischen Nationalbibliothek Tutorials zu Annif, da das Toolkit in immer mehr Einrichtungen

1 https://www.b-u-b.de/archiv/pdf-archiv-bub/pdf-archiv-detailseite/06/2022%20Schwerpunkt:%20Künstliche%20Intelligenz, Seite 306–311

2 Die genaue Zahl lässt sich über den Suchschlüssel ‚has:subject_stw_added' im Rechercheportal der ZBW (www.econbiz.de) ermitteln. Bisher werden außer Englisch, was ca. die Hälfte des Bestandes abdeckt, keine weiteren Publikationssprachen verarbeitet.

3 https://www.da-3.de/

4 Aktuell ist etwa die Integration des Deep-Learning-Modells X-Transformer (github.com/OctoberChang/X-Transformer) geplant – zu Transformermodellen siehe auch Kapitel 2.

auf Interesse stößt und auf sein Einsatzpotenzial geprüft wird.

Ein zentrales Thema ist die Qualitätskontrolle. Hier kommt ein umfassendes Konzept mit verschiedenen Ansätzen zum Tragen: Auf der technischen Seite werden dem Verschlagwortungsvorgang mehrere Filter mit zu erfüllenden Schwellwerten und diversen Blacklists zum Ausschluss bekannter problematischer Konstellationen nachgeschaltet. Ein bedeutender Teil der Qualitätssicherung beruht jedoch auf einer guten Zusammenarbeit mit den Erschließenden der ZBW – bisher sind die AutoSE-Modelle allein auf Verschlagwortungen aus deren täglicher intellektueller Arbeit trainiert, sie wirken fortlaufend an der Weiterentwicklung des hierbei verwendeten Standard-Thesaurus Wirtschaft (STW)[5] mit und im Laufe der Pilotphase bewerteten sie regelmäßig Stichproben des automatisiert generierten Outputs, was wertvolle Hinweise auf systematische Fehler lieferte. Seit 2022 geben die Erschließenden im DA-3 durchgängig für jeden von ihnen bearbeiteten Metadatensatz, der AutoSE-Vorschläge enthält, über ein integriertes Tool auch eine Bewertung dieser Vorschläge ab. So kann das Team Erkenntnisse darüber gewinnen, wie der Output des Dienstes angenommen wird, bekommt Hinweise auf schwer zu verarbeitende Publikationen und kann mit diesen Zusatzinformationen die eingesetzten Methoden weiter verbessern.

Eine Kernkomponente des Dienstes ist das von der Finnischen Nationalbibliothek entwickelte Open-Source-Toolkit Annif, welches verschiedene Standard-Algorithmen für eine automatisierte Inhaltserschließung anbietet.

Aus der Anfang 2024 offiziell zu Ende gegangenen Pilotphase lassen sich mehrere Zwischenfazits ziehen:[6] Zunächst einmal benötigt das Überführen von prototypischen Ansätzen aus der angewandten Forschung in tatsächlich benutzbare Praxislösungen und das anschließende Aufrechterhalten eines Produktivbetriebs grundsätzlich mehr Ressourcen (qualifiziertes Personal, Soft- und Hardware, aber auch Aufmerksamkeit!), als allgemein angenommen wird. Ein essenzieller Schritt für AutoSE war der Statuswechsel 2019 vom Projekt zur Daueraufgabe, um der Tatsache Rechnung zu tragen, dass die Automatisierung der Inhaltserschließung Bibliotheken noch viele Jahre begleiten wird – das stärkte das Commitment aller Beteiligten im Haus und führte konkret über die Symbolik hinaus zur Einrichtung einer dringend benötigten weiteren Stelle. Da es regalfertige maschinelle Erschließungssysteme (aktuell noch) nicht gibt, müssen verfügbare Open-Source-Lösungen mit verschiedenen Expertisen begleitet und angepasst werden, und alle Rollen müssen besetzt sein: die Forschungsarbeit im KI-Bereich, die Softwareentwicklung, das Monitoring und die Administration – und zwar aufgrund der für diese Domäne typischen hohen Personalwechselfrequenz am besten doppelt. Auch nach Abschluss der Pilotphase gibt es im Produktivbetrieb quasi wöchentlich Brände zu löschen (kritische Updates, Reagieren auf Systemwarnungen, Umgang mit externen Veränderungen, …) und entsprechend benötigt ein langfristig angelegtes Automatisierungskonzept ein sorgfältig ausgearbeitetes Betriebsmodell. Und schließlich: Zur Sicherung der Qualität und zur Förderung der Akzeptanz empfiehlt sich eine enge Zusammenarbeit mit den (menschlichen) Erschließungsfachkräften.

2. Large Language Models ante portas – Bibliothekswesen, quo vadis?

Im Jahr 2018 stellte Google mit BERT eine neue Art von Sprachmodell vor, basierend auf einer sogenannten Transformerarchitektur (das »T« in BERT steht für »Transformers«), ungefähr zur selben Zeit begann OpenAI mit der Entwicklung seiner GPT-Sprachmodelle, und 2022 ging der auf GPT aufsetzende Chatbot ChatGPT an den Start und erfreut sich seither mit anderen, ähnlichen Anwendungen nicht enden wollender Aufmerksamkeit in der breiten Öffentlichkeit.[7] Auch in der Welt der Informationsinfrastruktureinrichtungen hat diese Disruption große Wellen geschlagen, und die Reaktionen reichen von begeisterten Zuschreibungen eines schier endlosen Potenzials für die Zunft bis hin zu einer Reihe von Ängsten und Sorgen – unter anderem deshalb, weil sie an den letzten Resten ihres einstmaligen Monopols in Bezug auf die Grundaufgabe zu rütteln scheint, die Informationsbedürfnisse von Nutzenden auch im Kontext komplexerer fachlicher Recherchen gezielt zu bedienen. Anwendungen auf der Basis sogenannter Large Language Models (LLMs) scheinen Perspektiven auf neue Formen der Suche zu eröffnen, bei der Nutzende ihre Fachfragen in natürlicher Sprache stellen können und eine ausformulierte Antwort zurückbekommen, statt sich durch eine Trefferliste von Literaturressourcen hindurcharbeiten zu müssen. Entsprechend stand in den letzten Jahren häufiger die Prognose im Raum, dass die erfolgreiche Umsetzung einer solchen Interaktion zwischen Nutzenden und Volltexten über LLMs sowohl die Inhaltserschließung als auch die gezielte Metadatenerstellung an sich – bisher Kerntätigkeiten von Bibliotheken – obsolet machen werde. Hier drängen sich natürlich eine Reihe weiterer Fragen auf: Wie wahrscheinlich ist eine solche Entwicklung, wie weit sind wir von einer erfolgreichen Umsetzung entfernt und was

5 www.zbw.eu/de/ueber-uns/wissensorganisation/standard-thesaurus-wirtschaft/. Siehe auch Andreas Oskar Kempf, Joachim Neubert: STW Thesaurus for Economics. In: Birger Hjørland, Claudio Gnoli (Hrsg.): Encyclopedia of Knowledge Organization, Edmonton. International Society of Knowledge Organization, 2021. hdl.handle.net/11108/471.

6 Anna Kasprzik: Automating subject indexing at ZBW – making research results stick in practice. In: LIBER Quarterly – The Journal of the Association of European Research Libraries 33(2023)1. doi.org/10.53377/lq.13579.

7 ai.v-gar.de/ml/transformer/timeline/

wären die Voraussetzungen dafür? Und: Welche Rolle kommt in all dem in Zukunft den Bibliotheken zu?

Zunächst einmal ist es mittlerweile Allgemeinwissen, dass sich einer LLM-basierten Anwendung durch geschickteres Formulieren der Eingabe (»Prompten«) passgenauere Ausgaben entlocken lassen – nicht umsonst ist Prompt Engineering heutzutage ein Berufsbild. Das bedeutet, für eine fruchtbare Interaktion ist weiterhin ein Verständnis dafür erforderlich, wie man ein Anliegen möglichst gut strukturiert in eine Anfrage übersetzt – die aus Suchportalen bekannte Expertisesuche hat insofern lediglich die Gestalt gewechselt. Ein größeres Problem für potenzielle Suchanwendungen auf der Basis der neusten Entwicklungen ist das folgende: Anwendungen wie ChatGPT waren bisher nur mit LLMs untersetzt, die aufgrund ihres Trainings auf riesigen Datenmengen aus dem Internet mit einer plausibel klingenden Collage von Formulierungen zu einem bestimmten Thema reagieren. Weil das Ergebnis häufig mit dem Kenntnisstand der Nutzenden übereinstimmt, scheinen viele die irrige Erwartung zu hegen, dass LLMs einen Mechanismus zur Prüfung dieses Outputs auf Konsistenz (in sich und mit bekannten Informationen) enthalten, und der Anwendung »Halluzinationen« zuzuschreiben, wenn das einmal nicht der Fall ist – so sind LLMs aber nicht konzipiert. Allerdings: Dass ein Abgleich mit etabliertem Wissen ein großer Fortschritt wäre, ist unbestritten.

Nicht nur das Training, sondern auch der Betrieb von Modellen wie GPT verbraucht extreme Mengen von Ressourcen,[9] während eine Nachschlageoperation in einer Wissensbasis um ein Vielfaches sparsamer abläuft.

An diesem Punkt setzt eine Strategie namens Retrieval-Augmented Generation (RAG)[8] an, bei der die Anwendung zunächst eine vorab festgelegte Wissensbasis konsultiert und ihre Antwort dann auf der Basis der dort gefundenen Informationen formuliert. Sowohl bei der Vorverarbeitung der Quellen für die Wissensbasis als auch beim Abgleich der Anfrage mit den vorhandenen Informationen ist ein Element zentral: Metadaten. Bibliotheken könnten also weiterhin eine Rolle spielen, indem sie sich auf Metadatenerstellung als Kernkompetenz konzentrieren, diese aber in für LLM-Anwendungen verarbeitbare Formen (aktuell zum Beispiel Vektorbasen oder Wissensgraphen) gießen und zur Nachnutzung für Forschung, Entwicklung und Betrieb bereitstellen – der hierfür erforderliche Innovationsschub müsste allerdings angesichts der in Bibliotheken weiterhin flächendeckend genutzten Altformate gezielt und zügig erfolgen. Größere Kompetenzzentren könnten unterdessen daran arbeiten, RAG-basierte Suchportale anhand der Prinzipien von öffentlich geförderten Informationsinfrastrukturen zu entwickeln, sodass diese ihrem Auftrag als Dienstleister für die Wissenschaft auch in Zukunft gerecht werden und die neuen Technologien ohne kommerzielle Interessen einem größtmöglichen Nutzen für Forschende und für die Allgemeinheit zuführen können.

Neben dem gewünschten Effekt der Untersetzung mit kontrollierbaren Informationen hat der RAG-Ansatz übrigens einen weiteren großen Vorteil: Nicht nur das Training, sondern auch der Betrieb von Modellen wie GPT verbraucht extreme Mengen von Ressourcen,[9] während eine Nachschlageoperation in einer Wissensbasis um ein Vielfaches sparsamer abläuft. Entsprechend sollten Einrichtungen, die sich Nachhaltigkeit als einen zentralen Wert auf die Fahne schreiben, genau abwägen, welche Einsatzszenarien von LLM-Anwendungen eine derartige Verschärfung der sich ohnehin beschleunigenden Klimakatastrophe überhaupt rechtfertigen – und möglichst auf offene, ressourcenschonendere Modelle und Kooperationen mit anderen Institutionen setzen.

Anmerkung: Dieser Text wurde von der verfassenden Person gänzlich ohne Einsatz von generativer Künstlicher Intelligenz erstellt.

3. Ausblick auf die zukünftige angewandte Forschung im AutoSE-Kontext an der ZBW

Es gehört zum Konzept von AutoSE, dass die in den Dienst integrierten Machine-Learning-Methoden und die Komponenten der Softwarearchitektur auch über die Pilotphase hinaus kontinuierlich auf den neusten wissenschaftlichen und technologischen Stand angehoben werden. So hat das Team nun Transformermodelle als jüngsten Ansatz aus dem Deep Learning aufgegriffen und untersucht sie spezifisch auf ihre Praxistauglichkeit für den Anwendungsfall der Inhaltserschließung. Da sie intern mit sprachunabhängigen Datenrepräsentationen arbeiten, bieten sich diese besonders für eine multilinguale Erschließung an. In einer ersten Experimentalphase werden

8 https://neo4j.com/blog/what-is-retrieval-augmented-generation-rag/

9 Siehe zum Beispiel www.brusselstimes.com/world-all-news/1042696/chatgpt-consumes-25-times-more-energy-than-google oder www.scientificamerican.com/article/what-do-googles-ai-answers-cost-the-environment/

10 Lakshmi Rajendram Bashyam und Ralf Krestel: Advancing Automatic Subject Indexing: Combining Weak Supervision with Extreme Multi-label Classification. In: Proceedings of the 1st International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP) collocated with the Extended Semantic Web Conference (ESWC). Springer, 2024. www.ipr.informatik.uni-kiel.de/publications/pdfs/nslp24.pdf.

vortrainierte LLMs in Annif eingebunden und an die Metadatenbestände der ZBW angepasst (»Finetuning«), um herauszufinden, welche Leistung sie im direkten Vergleich zu den bisherigen Modellen erbringen. Vorläufige Ergebnisse scheinen die Einschätzung zu bestätigen, dass LLMs alleine aufgrund der an Bibliotheken typischen Datenlage – eher kleine und heterogene Trainingsdatenmengen – voraussichtlich keinen Quantensprung in der Performanz erzielen werden.[10] Entsprechend wird das Team mit weiteren Ansätzen experimentieren, etwa mit einem RAG-ähnlichen Ansatz, bei dem der STW als zusätzliche Quelle von Kontextinformationen dient. Darüber hinaus hat die Tatsache, dass nun mit dem Auto-SE-Dienst bereits eine Automatisierungsinfrastruktur besteht, einen niederschwelligeren Einstieg dafür geschaffen, die Inhaltserschließungspraxis an der ZBW Stück für Stück weiter zu transformieren, etwa durch ein Verschieben der Rollen von Mensch und Maschine hin zu einem Ansatz namens Active Learning, bei dem die Maschine Menschen interaktiv gezielt um die Annotation eines Datensets ersuchen kann.

Die Expertise der Erschließenden im Bereich der Wissensorganisation wird also weiterhin gebraucht, fließt aber über neue und anfangs sicher ungewohnte Kanäle in den Erschließungsvorgang ein – ein solcher Wandel erfordert Offenheit und Mut und kann entsprechend nur durch eine vertrauensvolle Zusammenarbeit mit allen Beteiligten gestemmt werden.

Dr. Anna Kasprzik (na, nas) koordiniert seit 2019 die Automatisierung der Sacherschließung (AutoSE) an der ZBW – Leibniz-Informationszentrum Wirtschaft. Na hat Linguistik, Informatik und Psychologie studiert und 2012 im Bereich der Theoretischen Informatik promoviert. Nach einem Bibliotheksreferendariat am KIM Konstanz folgten Tätigkeiten in einem IT-Projekt beim Bibliotheksverbund Bayern und in der Forschung und Entwicklung an der TIB Hannover. Nas Interessen umfassen verschiedene Ansätze aus der Künstlichen Intelligenz, insbesondere semantische Technologien, Machine-Learning-Verfahren und die Frage, wie diese beiden Herangehensweisen für Inhaltserschließung und Retrieval sinnvoll verzahnt werden können.

Anmerkung: Anna Kasprzik ordnet sich nicht in das binäre Geschlechtersystem ein und entsprechend wird in der Vita die zweite Silbe des Vornamens »na« als alternatives Neopronomen anstatt »er« oder »sie« genutzt.