Literatur schneller erschließen dank Künstlicher Intelligenz

Künstliche Intelligenz ist seit vielen hundert Jahren ein Menschheitstraum. Gut bekannt sind die Geschichten rund um den Homunculus, den Golem oder Frankensteins Monster. Seit etwa einem Jahrzehnt ist Künstliche Intelligenz nun keine Science Fiction oder Laboridee mehr, sondern Teil unseres Alltags. Wir alle kennen Siri, Alexa, DeepL, autonome Fahrzeuge oder Drohnen. In den digitalen Medien fällt es uns schwer, Bots von Menschen zu unterscheiden. KI kann Entscheidungen und Aktionen übernehmen, die bisher Menschen getroffen und ausgeführt haben. Auch in der ZBW spielen KI-Methoden eine bedeutende Rolle für den digitalen Wandel von Geschäftsabläufen. Ein Praxisbericht.

KI-Methoden in den ZBW-Alltag integrieren
Wissenschaftliche Bibliotheken arbeiten derzeit in einem Wissenschaftsbetrieb, in dem laut UNESCO-Wissenschaftsbericht weltweit etwa 8 Millionen Forscher*innen publizieren. Pro Jahr steigt die Zahl wissenschaftlicher Aufsätze um etwa 8 Prozent. Nun ist es die Kernaufgabe von Bibliotheken, dafür zu sorgen, dass die Nutzer*innen sich in der Menge dieser Veröffentlichungen zurechtfinden und schnell und einfach Literatur zu einem bestimmten Thema finden. Dazu werden Textdokumente mit Metadaten beschrieben. Dieses Beschreiben von Aufsätzen, Monografien, Sammelbänden usw. passiert zum einen auf formaler Ebene (Autor*in, Titel, Seitenzahl usw.) und zum anderen auf inhaltlicher Ebene mit einem genormten Wortschatz. Geht es in dem Dokument beispielsweise um multinationales Marketing und internationalen Absatz oder um betriebliche Liquidität und Cashflow?

Allein aus den Wirtschaftswissenschaften erreichen die ZBW jährlich über 100.000 Titel, die auffindbar gemacht werden müssen. Zusätzlich entstehen mit dem digitalen Wandel zahlreiche neue Aufgaben für die Beschäftigten der ZBW, die ebenso erfüllt werden wollen. Daher müssen Automatisierungsstrategien entwickelt und eingesetzt werden, um auch weiterhin möglichst viele Publikationen inhaltlich mit der angemessenen fachlichen Tiefe zu erschließen.

Hier bieten Methoden aus der KI, konkret maschinelles Lernen, großes Potenzial für Erleichterungen bei der Arbeit.
Dr. Anna Kasprzik, Leiterin der Arbeitsgruppe Automatisierung der Sacherschließung, erläutert: „Erste Schritte in der Sacherschließung mit Methoden des maschinellen Lernens zu gehen ist einfach. Diese Methoden dann aber auch auf einen Stand zu bekommen, der unseren hohen Qualitätsansprüchen genügt, ist eine sehr komplexe Aufgabe, und mit der Entwicklung eines Prototypen ist sie noch lange nicht erfüllt. Unser Ziel ist es, Automatisierungslösungen über das Prototypstadium hinaus langfristig in die tägliche bibliothekarische Praxis einzubringen."

Trainingsdaten sind das A und O
Um nun Automatisierungslösungen nachhaltig in die täglichen Arbeitsabläufe bringen zu können, muss Vorarbeit geleistet werden. Eine grundlegende Voraussetzung sind geeignete Trainingsdaten, die der Maschine als Vorbild dienen können. Das sind derzeit die Metadatensätze, die von Expert*innen der ZBW erstellt wurden. Diese Trainingsdaten müssen in ausreichender Menge und in ausreichender Vielfalt vorliegen.

Als Textmaterial, aus dem die trainierte Maschine dann ihre Vorschläge für STW-Deskriptoren ableitet, werden aktuell nur sehr kurze Textpassagen verwendet, etwa der Titel oder von Autor*innen vergebene Keywords, weil man annehmen kann, dass diese die Publikation möglichst prägnant zusammenfassen, ähnlich wie ein Abstract. Damit die maschinelle Methode überhaupt einsetzbar ist, sollten allerdings möglichst viele solche Keywords in einem klar definierten Feld in den Metadatensätzen abgelegt sein.

Zudem benötigt die Arbeitsgruppe gut strukturierte rechtliche Informationen, die auch maschinenlesbar sind, um ihre Algorithmen korrekt im Rahmen der geltenden Text-und-Data-Mining-Rechte anwenden zu können. Saubere Metadaten sind also nicht nur beim Training wichtig, sondern auch in allen Szenarien, für die die Maschine trainiert worden ist.

Open Science im Kontext Automatisierte Sacherschließung
Auf dem Weg, die automatisierte Sacherschließung in die Praxis zu bringen, hat sich die ZBW 2019 unter anderem eng mit der Deutschen Nationalbibliothek und der Nationalbibliothek Finnland ausgetauscht. Insbesondere hat die Arbeitsgruppe die finnische Plattform Annif getestet und auf der Fachkonferenz SWIB – Semantic Web in Libraries wurde ein internationaler Anwenderworkshop veranstaltet. Ein wichtiger Schritt war es, das hierfür von der ZBW gestellte Trainingsmaterial zusammen mit Übungen zum Selbststudium als Open Data auf GitHub zu veröffentlichen (URL: https://github.com/NatLibFi/Annif-tutorial)

Über die bereits eingesetzten statistischen Verfahren hinaus evaluiert die Arbeitsgruppe ständig aktuelle Ergebnisse aus dem Gebiet der Künstlichen Intelligenz, etwa aus dem Deep Learning. Derzeit wird mit neuronalen Netzen experimentiert, um neue Verfahren zu entwickeln oder auch im Rahmen des bislang verwendeten Fusion-Ansatzes das Zusammenspiel der einzelnen Verfahren zu optimieren. Zudem ist es ein Ziel, die Qualitätsabschätzung zu automatisieren, so dass die noch nicht erschlossenen Dokumente jeweils dem am besten für sie geeigneten (maschinellen oder intellektuellen) Verfahren zugeführt werden können.

Im Jahr 2019 hat die ZBW 131.289 Publikationen formal erschlossen. Intellektuell konnten 26.372 Publikationen inhaltlich erschlossen werden. Über eine automatisierte Lösung wurden 2019 zusätzlich 86.396 Publikationen inhaltlich erschlossen.

INHALTSERSCHLIESSUNG

Bei der Inhaltserschließung (auch Sacherschließung genannt) arbeitet die ZBW mit einem genormten Wortschatz. Der sogenannte „Standard-Thesaurus Wirtschaft" (STW) bildet das weltweit umfassendste Fachvokabular zur Recherche und Erschließung wirtschaftswissenschaftlicher Inhalte. Verfügbar in Deutsch und Englisch, umfasst der STW knapp 6.000 Deskriptoren und über 20.000 Synonymverweise. Der STW wird in der ZBW von einem interdisziplinären Expertenteam inhaltlich an den aktuellen Sprachgebrauch in den Wirtschaftswissenschaften angepasst, kontinuierlich weiterentwickelt und technisch in vernetzte Informationsumgebungen und innovative Web Services eingebunden. Als einer der ersten Thesauri in der Linked Open Data Cloud veröffentlicht, enthält der STW auch Mappings zur GND, dem AGROVOC und zur JEL. Der STW ist unter der freien Datenbank-Lizenz ODbL (Open Database License) 1.0 verfügbar.