KI-gestützte Automatisierung der Sacherschließung mit Annif

ein Pilotprojekt

Alexandra Engel, Ralph Hafner

Die Sacherschließung ist eine der Kernaufgaben des Fachreferats in wissenschaftlichen Bibliotheken. Sie ist unerlässlich für eine effektive Recherche und den Zugang zu Informationen. Intellektuelle Inhaltserschließung ist zeitaufwendig, insbesondere, wenn man – wie wir am KIM – nach einer individuellen Bibliothekssystematik („Konstanzer Systematik“ kurz: „KonSys“) erschließt, die (noch) nicht mit anderen Systematiken und verbalen Inhaltserschließungssystemen vernetzt ist. Daher ist eine unmittelbare Nachnutzung der Inhaltserschließung anderer Bibliotheken nicht möglich. Die Erweiterung des Aufgabenspektrums um forschungsnahe Dienstleistungen, die zusätzlich zu den traditionellen bibliothekarischen Aufgaben im Fachreferat erbracht werden, bedeutet u.a., dass in den Fachreferaten weniger Zeit für die Sacherschließung zur Verfügung steht. Eine inhaltliche Erschließung der in den großen E-Book-Paketen enthaltenen Titel ist aufgrund ihrer Menge intellektuell nicht zu leisten. Ohne inhaltliche Erschließung sind diese Titel für unsere Nutzer*innen aber schlechter auffindbar.

Um diesen Herausforderungen zu begegnen, soll die Sacherschließung am KIM (teil-) automatisiert werden. Zum Einsatz kommen soll dabei die für automatisierte Sacherschließung konzipierte KI-Toolbox Annif. In einem Pilotprojekt sollte herausgefunden werden, ob die andernorts bereits erfolgreich eingesetzte Software Annif auch mit unseren Konstanzer Systematikdaten gute Ergebnisse liefert. Ziel des Pilotprojektes war es, anhand der bisher geleisteten intellektuellen Erschließung des Konstanzer Bestands mithilfe von Annif automatisiert Notationen für neu in den Bestand aufgenommene Titel zu generieren. Diese Notationsvorschläge sollen die Erschließungsarbeit bei Buchneuzugängen erleichtern oder im Falle von nicht erschlossenen E-Books direkt übernommen werden. Das Pilotprojekt, welches von Februar 2024 bis September 2024 durchgeführt wurde, diente primär dazu, erste Erfahrungen mit den KI-gestützten Modellen in Annif zu sammeln und diese mit Konstanzer Daten zu trainieren. Im Folgenden werden die Vorarbeiten als Voraussetzung des Projekts skizziert sowie das Tool Annif und die Ergebnisse des Pilotprojekts vorgestellt.

Vorarbeiten

Die ersten Schritte in Richtung Automatisierung der Sacherschließung wurden in den letzten Jahren erfolgreich abgeschlossen, mit dem Ergebnis, dass die Konstanzer Systematikdaten jetzt in interoperabler Form als Linked Open Data zur Verfügung stehen. Dazu mussten die hierarchischen Strukturen teils ergänzt, teils korrigiert und in eine maschinell auswertbare Form überführt werden. Die Systemstellenbenennungen wurden möglichst auf Begriffe der Normdatei GND gemappt und somit von ambivalenten Zeichenfolgen in eindeutige Konzepte mit IDs aus einer Normdatei überführt. (Hafner und Schelling 2015; Hafner 2022). Ohne diese umfassende Entwicklungs- und Bereinigungsarbeit an der Konstanzer Systematik wäre die Automatisierung und Vernetzung mit anderen Systematiken nicht realisierbar.

Pilotprojekt

Das hier vorgestellte Pilotprojekt untersuchte die Möglichkeit der automatisierten Sacherschließung nach der Konstanzer Systematik mithilfe der KI-Toolbox Annif. Annif ist ein open-source-Tool zur automatisierten Sacherschließung, das an der Finnischen Nationalbibliothek 2017-2020 entwickelt und gemeinsam mit der ZBW – Leibniz-Informationszentrum Wirtschaft weiterentwickelt wird (Suominen 2019; Suominen, Lehtinen, und Inkinen 2022). Das Softwarepaket ist durch seine modulare Architektur flexibel an die jeweiligen Erfordernisse vor Ort anpassbar. Es kann mit Vokabularen, Thesauri oder Systematiken „gefüttert“ werden. Zudem sind unterschiedliche Algorithmen des maschi-

nellen Lernens, sogenannte „Backends“, in Annif implementiert und können je nach Anwendungsfall ausgewählt werden. Im Analyzer-Modul wird die Sprache der zu erschließenden Titel, Volltexte und des Vokabulars festgelegt, sodass eine sprachspezifische Zerlegung und Verarbeitung der Texte mit computerlinguistischen Methoden möglich ist. Die REST-API ermöglicht die Verwendung einer webbasierten Eingabeoberfläche sowie die Integration in externe Systeme, wie beispielsweise den Digitalen Assistenten (DA3) oder Repositorien (Beckmann u. a. 2019). Annif wird in Deutschland bereits erfolgreich an der Deutschen Nationalbibliothek als Kern der Erschließungsmaschine EMa für verbale Sacherschließung mit der GND und für das Generieren von DDC-Kurznotationen eingesetzt (Mödden 2024). An der ZBW – Leibniz-Informationszentrum Wirtschaft wird Annif zur automatisierten Erschließung mit dem Standard-Thesaurus Wirtschaft (STW) eingesetzt (Kasprzik 2023).

Das Pilotprojekt umfasste die Aufbereitung der Daten sowie das Trainieren und Evaluieren erster Modelle. Zur Gewährleistung der Datenkonsistenz und zur einfacheren Evaluierung der Ergebnisse wurde eine Beschränkung auf den Systematikzweig der Linguistik vorgenommen. Die bibliographischen Daten wurden aus Excel-Dateien in ein für Annif lesbares tsv-Format gebracht und auf die Linguistik gefiltert (siehe Abbildung 1). Annif wertet den Text links des ersten Tabstopps aus und bringt die darin enthaltenen Begriffe in Zusammenhang mit den zugeordneten Notationen (im URI-Format). Im Beispiel in Abbildung 1 sind nur Autor und Titel enthalten, optional können auch Abstracts oder Inhaltsverzeichnisse links vom ersten Tabstopp stehen. Die Systematikdaten wurden in ein SKOS-Format überführt (siehe Abbildung 2), das auf dem Resource Description Framework (RDF) basiert und die hierarchische Struktur maschinenlesbar abbildet. Für das Training wurden die bibliographischen Daten in einen Trainings- und einen Testdatensatz (80% bzw. 20% der Gesamtmenge) geteilt. Die Modelle wurden auf die Titel im Trainingsdatensatz trainiert und deren Vorhersagen für die Titel im Testdatensatz evaluiert. Englisch- und deutschsprachige Modelle wurden getrennt trainiert. Die Modelle mit den Omikuji-Algorithmen (Bonsai, Parabel, Attention) erzielten die besten Ergebnisse im Vergleich mit der intellektuellen Erschließung.

	A	B	C	E	H	I	J
1	RID	MainAuthor	Title	ErschJahr	Language	ISBN	Alle_Notationen_MARC
2	323512283	Graffi, Giorgio	200 years of syntax	2001	eng	9027245878	spr 181.50,spr 8:t,spr 8:s
3	322220920	Bybee, Joan L.	Frequency and the emergence of linguistic structure	2001	eng	1588110273	spr 31.50
4	324109628	Aikhenvald, Alexandra Y.	Non-canonical marking of subjects and objects	2001	eng	1588110435	spr 197

Graffi, Giorgio 200 years of syntax → http://konsys.uni-konstanz.de/nt/spr%20181.50 → http://konsys.uni-konstanz.de/nt/spr%208%3At → http://konsys.uni-konstanz.de/nt/spr%208%3As

Bybee, Joan L. Frequency and the emergence of linguistic structure → http://konsys.uni-konstanz.de/nt/spr%2031.50

Aikhenvald, Alexandra Y. Non-canonical marking of subjects and objects → http://konsys.uni-konstanz.de/nt/spr%20197

Abbildung 1: Bibliographische Metadaten, transformiert von Excel nach tsv

Abbildung 2: Systematikdaten für die Systemstelle spr 12:b33 im SKOS-Format (rechts)

Abbildung 3 zeigt die Ergebnisse der automatisierten Sacherschließung mit Annif mithilfe des englischsprachigen Omikuji Bonsai-Modells für den Titel „A Typology of Reference Systems“ von Frajzyngier. Die intellektuelle Erschließung für diesen Titel ist spr 15.95:r33 Referenz <Linguistik>, spr 54.25 Grammatik (unter der Erstreckung Sprachtypologie) und spr 54.31 Syntax (ebenso Sprachtypologie). Die Qualität der automatischen Erschließung ist abhängig davon, welche Wörter in der Eingabe vorkommen. In diesem Fall sind die Begriffe typology und reference ausreichend für eine Zuordnung zu spr 15.95:r33 Referenz <Linguistik> und spr 53 Gesamtdarstellung (Sprachtypologie). Zusätzlich werden andere Systemstellen vorgeschlagen, an denen Titel mit ähnlichen Begriffen notiert sind oder die Verweisungen auf die GND-Begriffe haben (z. B. spr 233.50 Einzeldarstellung (Semantik)).

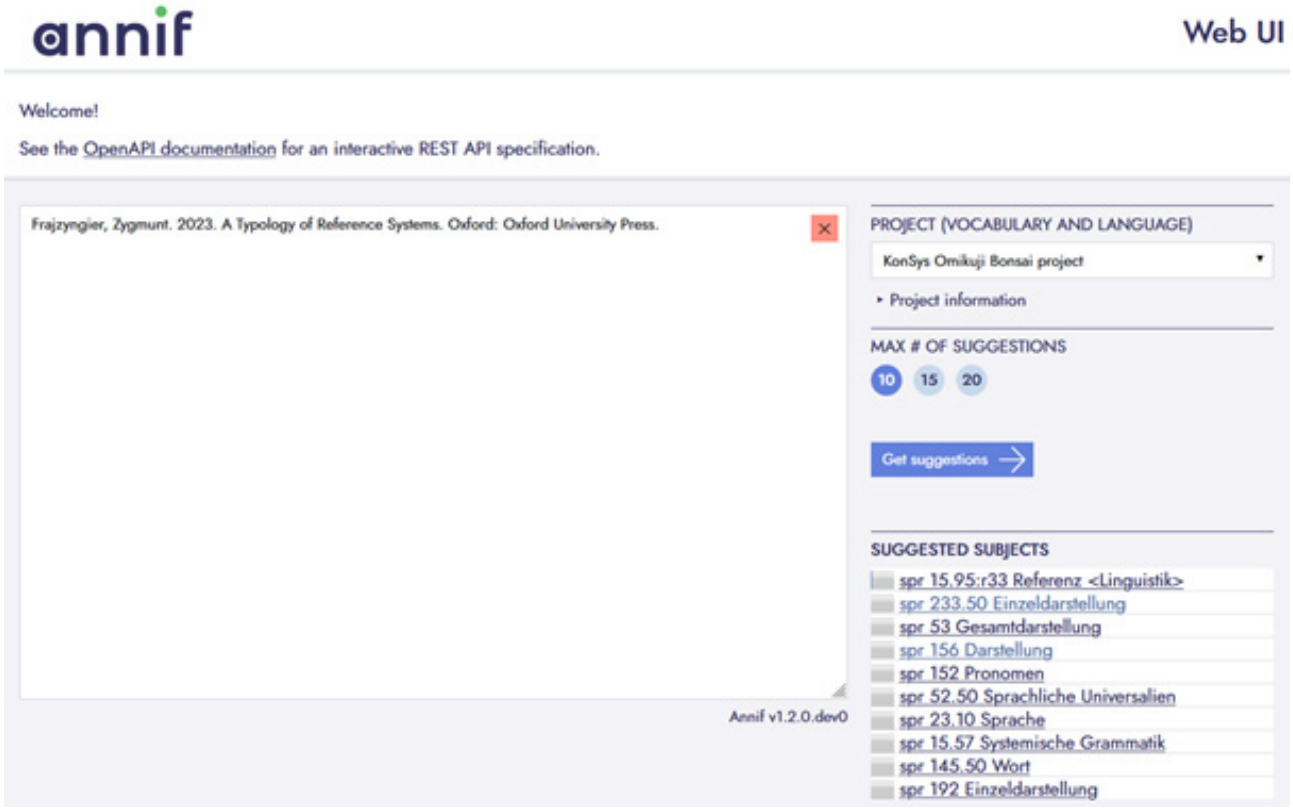

Abbildung 3: Ergebnisse der automatisierten Sacherschließung mit Annif

Ausblick

Das Pilotprojekt bietet eine gute Grundlage für die angestrebte Weiterentwicklung der automatisierten Sacherschließung mit Annif. Es zeigt, dass Annif grundsätzlich auch mit unserer komplexen Konstanzer Systematik zufriedenstellende Ergebnisse liefert, und das obwohl Feinjustierungen im System – aus Zeitgründen – noch nicht vorgenommen werden konnten. Zu diesen noch ausstehenden Feinjustierungen gehört u.a. die Hyperparameteroptimierung, was bedeutet, dass die Konfiguration der Modelle generisch und nicht auf die vorliegenden Daten abgestimmt war. Dieser Aspekt muss im weiteren Verlauf adressiert werden und beinhaltet umfassende Tests mit verschiedenen Modellkonfigurationen. Ein Problem stellt außerdem die „Longtail“-Charakteristik der Daten dar. Das bedeutet, dass an wenigen Systemstellen viele Titel und an vielen Systemstellen wenige Titel notiert sind. So sind an ca. 85% der Systemstellen weniger als 10 Titel notiert. Die Entwickler*innen von Annif empfehlen ca. 50 Titel für ein Schlagwort (in unserem Fall Systemstelle). Möglich wäre die Erschließungstiefe zu beschränken, sodass die Modelle nicht auf die unterste Ebene trainiert werden, sondern manche Systematikbereiche „zusammengefasst“ werden und somit mehr Daten pro Notation/Erstreckung vorhanden sind. Dies wird unter Umständen von Fach zu Fach unterschiedlich zu bewerten sein. Des Weiteren gilt es, Schwellenwerte für die Konfidenz von vorgeschlagenen Notationen festzulegen. Ab welchem Konfidenzwert wird die maschinelle Erschließung als qualitativ gut erachtet und ggf. direkt übernommen? Darüber hinaus wird auch die Möglichkeit von dynamischen Neural-Network-Ensembles zu prüfen sein, die bei

jeder Erschließungsaufgabe von der Eingabe und den generierten Notationen lernen und diese bei neuen Eingaben mitberücksichtigen.

Danksagung

Wir möchten Sorin Gheorghiu ganz herzlich für seine technische Unterstützung bei der Installation von Annif und der Bereitstellung der Systematikdaten im SKOS-Format danken. Damaris Schafranek danken wir für die Bereitstellung der bibliographischen Daten aus Libero. Und ich, Ralph Hafner, möchte mich ganz herzlich bei Dir, Alexandra, für Dein Engagement für dieses Pilotprojekt bedanken, das uns bei unserem Vorhaben, hier im KIM zu einer teilautomatisierten Sacherschließung zu kommen, ein großes Stück vorangebracht hat.

Literaturverzeichnis

Beckmann, Regine, Imma Hinrichs, Melanie Janßen, Gérard Milmeister, und Peter Schäuble. 2019. „Der Digitale Assistent DA-3 – eine Plattform für die Inhaltserschließung“. *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB* 2019 (3): 1-20. https://doi.org/10.5282/O-BIB/2019H3S1-20.
Hafner, Ralph. 2022. „Konsys. Das neue Tool für die Konstanzer Bibliothekssystematik“. KIM Kompakt 106:35-40.

Hafner, Ralph, und Bernd Schelling. 2015. „Automatisierung der Sacherschließung mit Semantic-Web-Technologie“. *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB*, Dezember, 161-175 Seiten. https://doi.org/10.5282/O-BIB/2015H4S161-175.

Kasprzik, Anna. 2023. „Aufbau eines produktiven Dienstes für die automatisierte Inhaltserschließung an der ZBW“. *o-bib. Das offene Bibliotheksjournal / Herausgeber VDB*, Februar, 1-13 Seiten. https://doi.org/10.5282/O-BIB/5903.

Mödden, Elisabeth. 2024. „Maschinelle Beschlagwortung mit Algorithmen“. *b.i.t. online* 27 (3): 242-53.

Suominen, Osma. 2019. „Annif: DIY automated subject indexing using multiple algorithms“. *LIBER Quarterly: The Journal of the Association of European Research Libraries* 29 (1): 1-25. https://doi.org/10.18352/lq.10285.

Suominen, Osma, Mona Lehtinen, und Juho Inkinen. 2022. „Annif and Finto AI : Developing and Implementing Automated Subject Indexing“. *JLIS*, Nr. 1. https://doi.org/10.4403/jlis.it-12740.