Matthias Nagelschmidt, Christa Schöning-Walter

Fachtagung Netzwerk maschinelle Verfahren in der Erschließung

Am 18. und 19. November 2021 hat die Deutsche Nationalbibliothek (DNB) zum dritten Mal die Fachtagung »Netzwerk maschinelle Verfahren in der Erschließung«¹ ausgerichtet, pandemiebedingt zum zweiten Mal virtuell. Themenschwerpunkt waren diesmal die Technologien der »Künstlichen Intelligenz« (KI). Damit knüpfte die Fachtagung an das Forschungsprojekt »Automatisches Erschließungssystem – Inhaltliche Erschließung von Publikationen mit KI« an, welches 2021 im Rahmen der nationalen KI-Strategie der Bundesregierung auf den Weg gebracht werden konnte. In diesem Projekt, das von der Beauftragten der Bundesregierung für Kultur und Medien (BKM) gefördert wird, will die DNB einige der schwierigen Herausforderungen der maschinellen Erschließung grundlegend neu angehen. Technologische Entwicklungen, die sich für die inhaltliche Erschließung textbasierter Medienwerke mit dem Vokabular der Gemeinsamen Normdatei (GND) eignen, sollen für die Erschließungsaufgaben der Bibliothek adaptiert werden. Der Informations- und Erfahrungsaustausch mit anderen forschenden, entwickelnden und anwendenden Institutionen spielt in diesem Zusammenhang eine wichtige Rolle. Das Netzwerk maschinelle Verfahren in der Erschließung ist eine Plattform für den Dialog. Mit 150 Teilnehmenden war die Fachtagung 2021 ausgebucht.

Die Reihe der Fachbeiträge wurde von Professor Kristian Kersting eröffnet, Lehrstuhlinhaber für KI und Maschinelles Lernen am Fachbereich Informatik der Technischen Universität Darmstadt. In einer perspektivisch ausgerichteten Keynote zu den rasanten Entwicklungen skizzierte er die Fähigkeiten der nächsten Generation KI-basierter Technologien, die anhand riesiger Datenmengen nicht nur mathematisch abgeleitete Vorhersagen treffen können, sondern bereits Formen abstrakten Denkens und logischen Schlussfolgern aufweisen. Kennzeichen dieser »dritten Welle der KI« ist ein Kreislauf von

kontextabhängigem Entscheiden und Lernen in sich verändernden Situationen.

Sven Giesselbach vom Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme (IAIS) stellte anschließend den Anwendungsbezug zum Arbeitsbereich Natural Language Understanding (NLU) her. Er zeigte sehr anschaulich, wie durch Kombination klassischer Methoden mit regelbasiertem Wissen logische Zusammenhänge erkannt und abgebildet werden. Interaktive Inhaltsverzeichnisse für Dokumentkollektionen oder die Bestimmung semantischer Ähnlichkeit einzelner Terme oder ganzer Textdokumente sind Anwendungsbeispiele dafür.

Professor Georg Rehm, der am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) im Bereich der Sprachtechnologien arbeitet, lenkte den Blick auf das Thema »digitale Sprachgerechtigkeit« und stellte das European Language Grid vor. In diesem Projekt der Europäischen Union entsteht eine zentrale europäische Plattform für Sprachtechnologien. Auch Kuratierungstechnologien, die in den letzten Jahren im Projekt QURATOR entstanden sind, sollen dort einfließen. Das sind mehr als hundert Services zur Extraktion, Analyse oder Anreicherung von Daten und Texten.

Die Datenqualität stand im Mittelpunkt der zwei anschließenden Beiträge. Lydia Pintscher, Produktmanagerin von Wikidata bei Wikimedia Deutschland, gab einen Einblick in das Qualitätsmanagement für diese freie und multilinguale Wissensdatenbank. Die Ursachen für fehlerhafte Daten, die meistens unbeabsichtigt entstehen, können vielfältig sein. Zielkonflikte oder Fehler bei der Datenmodellierung sowie sprachliche, kulturelle, thematische und technische Fehler sind Beispiele dafür. Lydia Pintscher zeigte verschiedene Ansätze, die Qualitätssicherung mit automatisierten Verfahren zu unterstützen und die Datenqualität messbar und sichtbar zu machen. Verknüpfungen und Vergleiche spielen dabei eine wichtige Rolle. Der Vor-

trag schloss mit dem Fazit, dass sich die Offenheit von Wikidata in diesem Zusammenhang sowohl vorteilhaft als auch nachteilig auswirkt.

Jürgen Kett, Leiter der Arbeitsstelle für Standardisierung der DNB, präsentierte entsprechende Überlegungen für die GND. Er machte die Vielschichtigkeit des Themas »Verlässlichkeit der Daten« deutlich und sprach Aspekte an, die von der Organisation und Infrastruktur über Standards und Regeln bis zu Expertise, Qualitätsbewusstsein und Akzeptanz reichten. Aufgrund der hohen Qualitätsanforderungen, die für Normdaten gelten müssen, ist die Nutzung automatisierter Verfahren für die Pflege der Normdaten besonders schwierig und deshalb bislang kaum etabliert. Wie kann die Eignung der GND für den Einsatz automatisierter Verfahren verbessert und die redaktionelle Arbeit unterstützt werden? Wie finden neue begriffliche Konzepte den Weg in die GND, wenn die Publikationen dazu ausschließlich maschinell erschlossen werden? Solche Fragen wurden aufgeworfen und diskutiert.

Die Bedeutung von Open Source im Bereich der Verarbeitung natürlichsprachlicher Informationen beleuchtete Stefan Geißler, Mitbegründer der Firma Kairntech. Der Software- und Dienstleistungsanbieter entwickelt Informations- und Retrievalsysteme und setzt dabei zunehmend auf Open-Source-Produkte. Stefan Geißler begründete dies mit den Vorteilen der hohen Innovationsdynamik, die die Open-Source-Bewegung auszeichnet. Zudem können Abhängigkeiten vermieden werden, die bei Anwendung proprietärer Technologien auftreten. Deshalb ist eine Mitwirkung in der Open-Source-Community auch für kommerzielle Unternehmen interessant, so sein Fazit.

Hans-Jörg Schäuble, in leitender Funktion beim Start-up-Unternehmen Aleph Alpha tätig, richtete den Blick zum Abschluss des ersten Tages noch einmal in die Zukunft. Er hob das enorme Potenzial der transformativen KI hervor und formulierte die Prognose, dass mehr als fünfzig Prozent der heute von Menschen verrichteten Tätigkeiten innerhalb der kommenden zwanzig Jahre durch KI lösbar sein werden. Das Sprachmodell »Generative Pretrained Transformer 3« (GPT-3), welches etwa 172 Milliarden Parameter umfasst, ist ein Beispiel für eine solche Technologie der nächsten Generati-

on. Das Unternehmen Aleph Alpha hat den Anspruch, auf diesem Sektor in Europa maßgeblich zum Fortschritt beizutragen.

Der zweite Veranstaltungstag begann mit einer Präsentation von Osma Suominen, Chefentwickler an der Finnischen Nationalbibliothek, der zunächst das Toolkit Annif und den Indexierungsservice Finto AI vorstellte. Das Toolkit Annif integriert leistungsfähige Textanalyseverfahren, einige der Module basieren auf Machine-Learning beziehungsweise Deep-Learning-Ansätzen. Auch Osma Suominen sprach Aspekte der Qualitätsmessung und -bewertung intellektueller und maschineller Erschließungsverfahren an und formulierte pointiert: »Die Fehler menschlicher Indexierer sind meistens nachvollziehbar, die Fehler der Algorithmen erscheinen oft dumm.« Und: »Algorithmen erzielen fast immer dann die besseren Ergebnisse, wenn sie mit anderen Algorithmen kombiniert werden.« Er schloss den Beitrag mit einem Plädoyer für Systemoffenheit und Zusammenarbeit in der Open Source-Community.

Zu den Institutionen, die bereits seit längerem in dieser Community aktiv mitwirken, gehört die ZBW – Leibniz-Informationszentrum Wirtschaft. Annif ist ein wesentlicher Bestandteil der Softwareentwicklungen an der ZBW, über deren Architektur Anna Kasprzik und Moritz Fürneisen berichteten. Einige der technischen und konzeptionellen Herausforderungen der Herstellung und Einführung eines Produktivsystems für die automatisierte Inhaltserschließung wurden exemplarisch vorgestellt. Auch in diesem Beitrag waren Fragen der Qualitätssicherung ein zentrales Thema: Algorithmen machen Fehler, wie können diese gefunden und überwunden werden?

Zum Abschluss stellte Sandro Uhlmann mit dem Projekt EMa – Erschließungsmaschine den Stand der Entwicklungen an der DNB vor. Die DNB baut zurzeit ein neues Softwaresystem auf, um die bisherige Erschließungssoftware abzulösen. Auch hier werden Komponenten von Annif integriert. Die Erneuerung der Software wirkt sich auch auf die Erschließungsprozesse aus. Die gesamte Verarbeitungskette von der Textbereitstellung und Spracherkennung bis hin zur Verzeichnung der Erschließungsergebnisse in die bibliografischen Datensätze ist von den Veränderungen betroffen.

An den beiden Veranstaltungstagen wurde damit ein breites Informationsspektrum geboten: vom Ausblick auf die nächste Generation KI-basierter Technologien über Methoden zur Verarbeitung und Analyse natürlicher Sprache und Fragestellungen des Qualitätsmanagements bis hin zu den An-

wendungsbeispielen in der Praxis. Wir danken allen Referent*innen für die vielfältigen und anregenden Einblicke in die Forschungs- und Projektarbeit und den Teilnehmenden für ihr Interesse und die engagierten Diskussionsbeiträge.

Anmerkungen

1 <<https://wiki.dnb.de/x/kmv3D>>