Ein Feuerwerk an Algorithmen und der Startschuss zur Bildung eines Kompetenznetzwerks für maschinelle Erschließung

Bericht zur Fachtagung „Netzwerk maschinelle Erschließung“ an der Deutschen Nationalbibliothek am 10. und 11. Oktober 2019

Am 10. und 11. Oktober 2019 trafen sich rund 100 Vertreterinnen und Vertreter aus Bibliothek, Wissenschaft und Wirtschaft an der Deutschen Nationalbibliothek (DNB) in Frankfurt am Main zu einer Fachtagung über das derzeitige Trend-Thema „maschinelle Erschließung“. Ziel der Veranstaltung war die „Betrachtung unterschiedlicher Anwendungsbereiche maschineller Textanalyse“[1] sowie die Initiation eines Dialogs zu Technologien für die maschinelle Textanalyse, Aufgabenstellungen, Erfahrungen und den Herausforderungen, die maschinelle Verfahren nach sich ziehen. Hintergrund ist der Auftrag des Standardisierungsausschusses an die DNB, regelmäßig einschlägige Tagungen durchzuführen, aus denen „perspektivisch ein Kompetenznetzwerk für die maschinelle Erschließung entsteh[t]“.

Ute Schwens, stellvertretende Generaldirektorin der DNB, eröffnete die Tagung. Dabei hob sie hervor, dass in den letzten Jahren die digitale Transformation ein maßgebliches Thema für Kulturinstitutionen geworden sei. Das stetige Anwachsen der Menge digitaler Publikationen zwang die DNB vor rund 10 Jahren, sich aus Kapazitätsgründen mit der maschinellen Erschließung zu beschäftigen. Mittlerweile betreibt die DNB ein lernendes Verfahren zur Klassifizierung mit DDC-Kurznotationen und ein computerlinguistisches Verfahren für die verbale Inhaltserschließung – mit dem Ziel, stetig mehr Erschließungsleistung anzubieten. Dabei setze die DNB auf eine fortlaufende Verbesserung und Erneuerung der Verfahren.

Elisabeth Mödden, an der DNB verantwortlich für automatische Erschließungsverfahren und Netzpublikationen, und Christa Schöning-Walter (DNB) führten durch das Tagungsprogramm.

Können Wissensgraphen die Kommunikation in der Wissenschaft verändern?

Den Auftakt in der ersten Session machte Markus Stocker, Umweltinformatiker und Leiter der Nachwuchsforschungsgruppe Knowledge Infrastructures an der TIB – Leibniz-Informationszentrum Technik und Naturwissenschaften und Universitätsbibliothek, mit einem Vortrag zur Frage „Können Wissensgraphen die Kommunikation in der Wissenschaft verändern?“.

Mit Blick auf den derzeitigen wissenschaftlichen Publikationsprozess stellte Stocker fest, dass sich – im Vergleich zu anderen Branchen, die durch die Digitalisierung bereits starke Veränderungen erfahren haben – bei der Art, wie Wissen kommuniziert wird, in den letzten vier Jahrhunderten keine wesentlichen Änderungen ergeben haben. Wesentliche Bestandteile der Wissenskommunikation

1 Fachtagung Netzwerk maschinelle Verfahren in der Erschließung,<https://wiki.dnb.de/x/GwfmC>, Stand: 26.10.2019. Über diese Seite können auch die Vortragsfolien von der Veranstaltung abgerufen werden.

seien Titel, Abstract und ein sich anschließender Textkörper. Enthaltene Daten wie Tabellen u.ä. seien für Maschinen nicht lesbar oder auswertbar. Abhilfe schaffe der Open Research Knowledge Graph[2] (ORKG) – ein Wissensgraph, der derzeit an der TIB entwickelt wird. Ziel des ORKG ist die Schaffung einer Publikationsinfrastruktur, die es den Autor/inn/en bereits während des Prozesses des wissenschaftlichen Schreibens ermöglicht, wissenschaftliche Information z.B. durch Anwendungen wie Jupyter Notebook maschinenlesbar in ihre Publikationen zu integrieren. Weiterhin können zur Publikation gehörende formale Metadaten gecrawlt, eine klassifikatorische Grobzuordnung getroffen und die Forschungsfrage strukturiert erfasst werden. Durch diese Veränderungen im Publikationsprozess ermögliche der ORKG im späteren Retrieval durch eine spezielle Abfragesprache für Forschungsfragen präzisere Antworten, die Anzeige von Publikationen mit ähnlichen Fragestellungen sowie die Visualisierung als Knowledge Graph.

Derzeit befindet sich die Entwicklung des ORKG in der Alpha-Phase; weitere Features wie Versionierung, Benutzermanagement, disziplinenspezifische Anpassungen sowie Tools zur Qualitätssicherung sind in Arbeit. Darüber hinaus stehen Kooperationen mit Verlagen, der Publikationssoftware Open Journal Systems und anderen Nachweissystemen wie OpenAIRE sowie die Schaffung von Schnittstellen zu Forschungsinformationssystemen und der European Open Science Cloud auf der Agenda.

In der anschließenden Diskussion wurde darauf hingewiesen, dass dieser Ansatz ein Gegenmodell zu den bisherigen Verfahren der maschinellen Erschließung darstelle, da er ganz stark auf den intellektuellen Beitrag der Autor/inn/en setze. Betont wurde außerdem die Notwendigkeit, diesen geeignete Ontologien, normierte Vokabulare etc. zur Verfügung zu stellen – hier könnten durchaus auch bibliothekarische Werkzeuge zur Anwendung kommen.

Digitalisate kuratieren mit KI – von unstrukturierten Daten zu strukturierten Inhalten

Anschließend referierte Clemens Neudecker, Forschungsreferent an der Staatsbibliothek zu Berlin (SBB) und Vertreter für die SBB am Forschungsprojekt QURATOR: Curation Technologies, über das derzeitige Projektvorhaben QURATOR, ein vom Bundesministerium für Bildung und Forschung (BMBF) gefördertes Projekt zur branchenübergreifenden Verbesserung von Kuratierungstätigkeiten und Generierung digitaler Inhalte durch Automatisierung.[3]

Neudecker berichtete über das Teilprojekt der SBB für „Automatisierte Kuratierungstechnologien für das digitalisierte kulturelle Erbe“[4] und stellte mit Bezug auf den vorherigen Vortrag heraus, dass die Herangehensweise von QURATOR in umgekehrter Richtung erfolge: Während der ORKG schon zu Beginn des Publikationsprozesses greife, beschäftige sich QURATOR@sbb mit Technologien im Postprocessing, da als Datengrundlage die digitalisierten Sammlungen der SBB dienen. Derzeit

2 Öffentliche Alpha-Version: <https://labs.tib.eu/orkg/>, Stand: 26.10.2019; Projekt auf GitLab: <https://gitlab.com/TIBHannover/orkg>, Stand: 26.10.2019.
3 Qurator Curation Technologies, <https://qurator.ai/projekt/>, Stand: 26.10.2019.
4 Vgl. Neudecker, Clemens: Vom Bücherspeicher zur Informationsinfrastruktur. Die Bibliothek erfindet sich neu, Qurator Curation Technologies, <https://qurator.ai/partner/staatsbibliothek-zu-berlin/>, Stand: 26.10.2019.

liegen rund 16.000 Digitalisate mit rund 5 Mio. Seiten vor, darüber hinaus im Zeitungsportal ZEFYS weitere 7 Mio. Seiten, von denen 3 Mio. mit Optical Character Recognition (OCR) behandelt wurden.

Neudecker rückte den aktuellen Hype um künstliche Intelligenz (KI) in ein anderes Licht, indem er sie in den Kontext gängiger Verfahren stellte: „KI ist Stochastik und Algebra". Die Liste der anschließend vorgestellten, im QURATOR-Projekt eingesetzten Verfahren reichte von der Metadatenanalyse und -validierung über OCR (für Fraktur inkl. der Ausgabe von Ligaturen in Unicode), automatisierte OCR-Nachkorrektur, Layout- und Strukturerkennung bis hin zur Named Entity Recognition (NER) via BERT[5] und Named-Entity-Disambiguation und -Linking mittels Wikidata und GND. Alle Entwicklungen werden auf GitHub frei zur Verfügung gestellt.[6] Als künftige Arbeitsfelder nannte Neudecker z.B. die Bildähnlichkeitssuche in digitalisierten Abbildungen. Das Publikum war sehr interessiert an Aufwänden und Kosten und erfuhr, dass dies kein billiges Projekt und mit Hardware „aus dem Geschäft" nicht zu bewerkstelligen sei.

Textanalyse im Kontext der Rundfunkanstalten. Forschung und Praxis

Im Anschluss sprach Jens Fisseler, Diplom-Informatiker und Mitarbeiter am Fraunhofer-Institut für Intelligente Analyse- und Informationssysteme IAIS, über die maschinelle Erschließung audiovisueller Materialien für Rundfunkanstalten.

Das Fraunhofer-Institut, so Fisseler, habe jahrzehntelange Erfahrung im Bereich maschinellen Lernens. Er berichtete über die dort entwickelte multimediale Mining-Plattform zur Inhaltsanalyse. Herausforderungen dabei seien die Transkription von gesprochenem Wort, die Erkennung, Identifikation und Disambiguierung unterschiedlicher Sprecherinnen und Sprecher, die Suche im gesprochenen Wort sowie die Keyword Extraction – jeweils auf individuelle Bedürfnisse der Kunden abgestimmt. Auch beim IAIS setzt man auf klassische Verfahren der NER, der statistischen Textanalyse mit TF-IDF und Word2Vec. Während TF-IDF statistisch signifikante Wörter eines Dokumentes mit ihrer Signifikanz in einem Dokumentenpool vergleicht, errechnet Word2Vec einen hochdimensionalen Vektorraum, in dem semantisch ähnliche Wörter nahe beieinanderliegen.[7] Fisseler schloss mit einem Ausblick auf künftige Entwicklungsvorhaben: Hier fielen die Schlagwörter NER-Disambiguierung und -Linking, Semantic Tagging, Topic Modelling sowie die Umstellung der Extrahierung von Keywords auf Keyphrasen. Aus dem Publikum kam während der Diskussion die Frage nach Ideen zur Qualitätssicherung. Fisseler antwortete, dass Qualität vor allem von der Güte der Trainingsdaten abhänge.

5 <https://github.com/qurator-spk/sbb_ner>, Stand: 28.10.2019. BERT steht für Bidirectional Encoder Representations from Transformers und ist ein von Google entwickelter Algorithmus im Spektrum des Natural Language Processing, vgl. <https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html>, Stand: 28.10.2019.

6 <https://github.com/qurator-spk>, Stand: 26.10.2019.

7 Vgl. dazu auch die Wikipedia-Einträge zu TF-IDF, <https://de.wikipedia.org/wiki/Tf-idf-Ma%C3%9F>, und zu Word-2vec, <https://en.wikipedia.org/wiki/Word2vec>, Stand: 26.10.2019.

Die Vermessung der Welt – Maschinelle Erschließung des deutschsprachigen WWW

Nach der Pause eröffnete Joachim Feist von der Firma mindUp Web + Intelligence GmbH die zweite Session mit einem Blick in die Angebote der freien Wirtschaft zur maschinellen Erschließung des deutschsprachigen World Wide Webs.

Zunächst sprach Feist über zwei Verfahrensfelder: Zum einen ging es um die zielgerichtete Verarbeitung von E-Mails inkl. Erkennung von Sprache, Intention, Tonalität und Thematik (z.B. Unterscheidung von Anfragen zu Produkten und Beschwerden). Zum anderen wurde kontextsensitive Produktwerbung thematisiert, die je nach dem Inhalt von Webseiten, Blogs u.Ä. passende Werbung einblenden soll. Dafür wird das deutschsprachige Netz gecrawlt und inhaltlich analysiert. Feist verglich dies mit der maschinellen Vergabe von Sachgruppen und GND-Schlagwörtern durch die DNB. Er betonte, dass das Kategoriensystem der eingesetzten Software nicht fix sei, sondern kundenbezogen entwickelt, überarbeitet und aktualisiert werde. Deshalb müssten gecrawlte Seiten auch regelmäßig nachklassiert werden. Als technisches Verfahren wird ein so genanntes Bayes-Netzwerk[8] eingesetzt. Als drittes Arbeitsfeld nannte Feist die Erstellung einer Domänendatenbank DACH, in welcher zu jeder Webpräsenz Informationen wie Thema, Reichweite etc. gesammelt werden. Hilfreich sei diese z.B. zum Auffinden neuer Online-Shops, aber auch zum Erkennen von Fake-Shops oder für die Erstellung eines Verzeichnisses aller Homepages von Künstlerinnen und Künstlern im DACH-Raum.

Patentsuche und -überwachung mit Methoden der Künstlichen Intelligenz in Theorie und Praxis

Gleich zwei Vorträge beschäftigten sich mit maschinellen Verfahren im Kontext von Patenten. Kornél Markó, Computerlinguist und geschäftsführender Gesellschafter der Firma Averbis, stellte im ersten Vortrag unterschiedliche, in seinem Unternehmen entwickelte Verfahren der maschinellen Inhaltserschließung näher vor. Im Anschluss daran referierte Jochen Spuck von der Baseler Firma EconSight über die praktische Anwendung des Verfahrens Averbis Supervised Learning in großen Textkorpora von Patenten.

Markó schilderte den Hintergrund von Patentanalysen: Jährlich erschienen ca. 250.000 neue Patente, die intellektuell kaum noch zu bewältigen seien. Expert/inn/en wollten daher nur noch relevante Patente begutachten und bewerten, ob diese unter „Freedom to operate"-Geschichtspunkten (FTO)[9] den Spielraum für neue Produktentwicklung böten. Abhilfe würden hier Algorithmen aus dem Bereich maschinellen Lernens schaffen. Im Schnelldurchlauf präsentierte Markó dann unterschiedliche Modelle des maschinellen Lernens und deren Funktionsweisen, u.A. Support Vector Machine, Deep Learning

8 Vgl. den Wikipedia-Eintrag, <https://de.wikipedia.org/wiki/Bayes-Klassifikator>, Stand: 26.10.2019.
9 „Freedom to operate" bedeutet die Möglichkeit der Produktentwicklung, ohne das geistige Eigentum anderer zu verletzen.

mit Word Embeddings und zuletzt Convolutional Neural Network[10]. Die Vorteile sieht er in enorm besseren Ergebnissen im Vergleich mit anderen automatisierten Diensten. Anschließend berichtete Jochen Spuck über den Einsatz der Algorithmen von Averbis bei Econsight, vor allem im Hinblick auf die Aufgaben des Auffindens ähnlicher Patente, der Gruppierung größerer Mengen an Texten mit ähnlichen Inhalten, dem Klassieren und Kategorisieren sowie dem Monitoring – gerade für die Lösung der schwierigsten Aufgabe der Patentrecherche, die FTO. Dabei geht es um das Auffinden aller ähnlichen Patente, um das Risiko eines neuen Produkts einschätzen zu können. Das von Averbis entwickelte Verfahren löse, so Spuck, genau dieses Problem, sodass die Patentrechercheur/inn/en nunmehr nur noch vergleichsweise kleine Mengen von Dokumenten zu sichten hätten.

Maschinelle Erschließungsverfahren vs. Volltextsuche

Der erste Tag wurde abgerundet durch einen Vortrag von Christian Wartena, der Stoff zum Nachdenken für die Diskussion im Knowledge Café an Tag 2 lieferte. Wartena befasst sich an der Hochschule Hannover (HsH) mit Sprach- und Wissensverarbeitung und koordiniert den Studiengang Informationsmanagement.

Laut Wartena müsse man sich bei der Frage „maschinelle Erschließung vs. Volltextsuche“ genau anschauen, wozu das jeweilige Verfahren eingesetzt wird und ob hier nicht Äpfel mit Birnen verglichen werden. Während eine verbale Inhaltserschließung eine Boolesche Suche auf Schlagwortbasis ermögliche (ein Dokument ist mit einem bestimmten Schlagwort versehen oder nicht), sei zwar auch das Retrieval insofern binär, als sich Suchende für das Dokument entscheiden oder nicht, aber nicht jedes Dokument, das die Suchanfrage erfülle, sei gleich relevant. Im Gegensatz dazu ermögliche ein Volltextindex ein probabilistisches Retrieval – wenn ein Dokument in den Suchergebnissen weit oben auftauche, sollte es auch mit hoher Wahrscheinlichkeit durch die Terme in der Anfrage adäquat zusammengefasst werden und damit für den Suchkontext relevant sein.

Als Nächstes analysierte Wartena fünf potenzielle Vorzüge der Inhaltserschließung: 1) Trennung von Haupt- und Nebensache, 2) Terminologische Kontrolle, 3) Facettierung, 4) Kontrolle und Korrektur, 5) Vorschaufunktion.

Die Trennung von Haupt- und Nebenaspekt stellte für Wartena keinen Vorteil dar, da auch maschinelle Verfahren letztendlich eine Trennung in Haupt- und Nebenaspekte vornähmen und zusätzlich durch die Inhaltserschließung Nebenaspekte unterdrückt würden, die in gewissen Suchkontexten relevant sein könnten.

Zur terminologischen Kontrolle fragte Wartena am Beispiel der GND mit ihren 210.000 Sachbegriffen provokant, ob man hier überhaupt noch von Kontrolle im Sinne einer klaren Differenzierung der Begriffe sprechen könne. Hierauf wurde in der sich anschließenden Diskussion aus dem Publikum

10 Vgl. dazu auch die Wikipedia-Einträge zu Support Vector Machine, <https://de.wikipedia.org/wiki/Support_Vector_Machine>, und zu Convolutional Neural Network<https://de.wikipedia.org/wiki/Convolutional_Neural_Network>, Stand: 8.11.2019.

entgegnet, dass die hohe Zahl der Sachbegriffe in der GND nicht auf überzogener Kleinteiligkeit beruhe, sondern auf dem Charakter eines Universalthesaurus, der z.B. zahlreiche Schlagwörter für Tier- und Pflanzenarten, chemische Elemente, Krankheiten, historische Ereignisse oder Produktnamen enthalte.

Wartena wies außerdem auf eine Studie der HsH hin, der zufolge z.B. im B3Kat[11] nur etwa die Hälfte der Ressourcen inhaltlich erschlossen sei, und wiederum etwa die Hälfte davon mit mehr als einem Vokabular oder Klassifikationssystem. Durch diese Mischung der Wissensorganisationssysteme gingen aus seiner Sicht die Vorteile einer Erschließung „irgendwo auf dem Weg vom Bibliothekar zum Endnutzer verloren". Zudem seien Synonymnormalisierung und -expansion in der Volltextsuche mittlerweile Standardfeatures der Suchmaschinentechnologie, und Synonyme könnten statt über Thesauri auch über maschinelle Verfahren (Word2Vec) identifiziert werden.

Ein tatsächlicher Vorteil einer Inhaltserschließung mit Hilfe der oberen Ebenen einer Klassifikation oder eines Thesaurus sei hingegen die Möglichkeit zur Facettierung und die Kombinierbarkeit mit der Volltextsuche; dies erfordere aber eine möglichst hohe Abdeckung durch maschinelle Verfahren mit intellektueller Kontrolle. Des Weiteren fungierten die durch Inhaltserschließung gewonnenen Schlagwörter bei der Anzeige im Discovery System als eine Art verdichteter Abstract ("Vorschau") für die jeweilige Ressource.

Die Frage, ob Inhaltserschließung eine positive Auswirkung auf das Ranking in Suchportalen haben könnte, verneinte Wartena zunächst, da mit der aktuellen Praxis die inhaltliche Verdichtung, die durch die Erschließung gewonnen wird, durch die Indexierung für ein Portal (im Sinne des Durchsuchbarmachens relevanter Textteile im Metadatensatz und des Volltexts) wieder verwässert werde. Zusätzlich könne eine heterogene Erschließungslage verzerrende Auswirkungen auf das Relevanzranking haben. In der Diskussion später wurde angemerkt, dass durchaus verschiedene Möglichkeiten zur Auswertung von kontrollierten Schlagwörtern für ein Ranking denkbar seien.

Im Ausblick regte Wartena dazu an, über den Einsatz maschineller Verfahren nachzudenken, um eine höhere Konsistenz zu erreichen und Ergebnisse aus einer maschinellen Verschlagwortung (Termgewichte etc.) im Retrieval nachnutzen zu können, so dass eine sinnvolle Verschränkung von maschineller Inhaltserschließung und Retrieval möglich wird. Maschinelle Verfahren eröffneten darüber hinaus weitere Möglichkeiten für eine noch bessere Vernetzung und Ausschöpfung von Informationsquellen, z.B. als Ergänzung zu Crosskonkordanzen.

Parallele Workshop-Angebote

Am zweiten Tag wurden vier Workshops bzw. Hands-on Labs parallel angeboten: So erhielt eine Teilnehmergruppe unter der Leitung von Ramon Leon Voges und Nico Wagner (beide DNB) eine praktische Einführung in die Programmierung mit Python. Die übrigen drei Angebote werden im Folgenden ausführlicher dargestellt.

11 https://www.bib-bvb.de/web/b3kat

Vom Text zum Inhalt

Der Workshop „Vom Text zum Inhalt“ wurde von Matthias Nagelschmidt und Christoper Poley von der DNB geleitet. In einem informativen Kurzvortag zu Beginn des Workshops[12] wurden Schwerpunkte der maschinellen Inhaltserkennung in komprimierter Form vermittelt. Bei der Texterkennung werden zuerst verschiedene Verfahren der Mustererkennung zur Strukturierung der Texte eingesetzt. Die Teilnehmenden lernten Tokenisierung, Part-of-Speech-Tagging (POS-Tagging) und Chunking kennen. Tokenisierung segmentiert je nach Konfiguration einen Text in z.B. Wörter oder Sätze. Beim POS-Tagging werden die Wortarten und die Satzzeichen unter Beachtung des Kontexts erkannt. Chunking ist ein Verfahren der Zerlegung in kleinere Sinneinheiten.

Daneben wurden drei Methodentypen zur maschinellen Verarbeitung natürlicher Sprache vorgestellt: linguistisch-lexikonbasierte, linguistisch-regelbasierte und statistische Methoden. Bei den lexikonbasierten Methoden werden lexikale Einträge als Grundlage benutzt, bei den regelbasierten werden u.a. Regeln zur Wortstammerkennung eingesetzt. Bei den statistischen Methoden werden Wörter gezählt und statistisch ausgewertet. Anhand bereits ausgewerteter Texte und deren Verteilungen können dann auch neue Texte inhaltlich erschlossen werden.

Die verschiedenen Methoden können auch miteinander kombiniert werden. So kann mit den linguistischen Methoden ein Text bereinigt werden, beispielsweise indem Terme auf Grund- und Stammformen zurückgeführt werden. Mit statistischen Verfahren werden dann Begriffe gezählt und anhand der Ergebnisse können Aussagen über den Text getroffen werden. Das Problem der Eigennamen und ihrer Erkennung (Named Entity Recognition) wurde in einem eigenen Punkt erörtert.

Als Werkzeuge für den Praxisteil wurden Python3 und das Python NLTK (Natural Language Toolkit[13]) eingesetzt. Die beiden Dozenten hatten ein Set an Übungstexten aus der Medizin vorbereitet. In einem Python-Skript mit Lücken bei einigen Teilaufgaben sollten die Teilnehmenden die Lücken mit eigenen Vorschlägen füllen, um z.B. den Text zu tokenisieren und die vorkommenden Substantive zu identifizieren.

Die Teilnehmenden haben den Workshop am Schluss als gelungen bewertet, der Vortrag hat übersichtlich auch komplizierte Sachverhalte erklärt. Der Übungsteil war zwar ohne Grundkenntnisse von Python schwierig zu bewältigen, aber Dank der Hilfestellung der Dozenten nahm jede Teilnehmerin und jeder Teilnehmer auch daraus etwas mit.

12 Alle Materialen zu den Workshops finden sich unter: <https://wiki.dnb.de/display/FNMVE/Materialien+zu+den+Workshops>, Stand 29.11.2019

13 Vgl. dazu: <https://www.nltk.org/>, Stand 29.11.2019.

Knowledge-Café: Inhaltserschließung im digitalen Zeitalter für die Wissenschaft von heute und morgen

Das Knowledge Café lud ein, zu verschiedenen Themen im Kontext der Automatisierung der Inhaltserschließung zu diskutieren. Die beiden übergeordneten Themenblöcke waren „Maschinelle Inhaltserschließungsverfahren versus Volltextsuche", in dem u.a. über die Inhalte des Vortrags von Christian Wartena diskutiert werden konnte, und „Inhaltserschließung – Wie gut ist gut genug?", der durch Kurzvorträge von Jan Maas (SUB Hamburg) und Michael Franke-Maier (FU Berlin) eingeleitet wurde. Im ersten Themenblock wurde parallel an mehreren Tischen über folgende Themen diskutiert:

- Mehrwert der Inhaltserschließung
- Zusammenspiel der intellektuellen und der automatisierten Sacherschließung
- Potenziale der semantischen Verknüpfung

Die Teilnehmenden wechselten dabei in vorgegebenen Zeitabständen die Tische, sodass alle Themen von allen Anwesenden diskutiert werden konnten.

In der breit gefächerten Diskussion zeichnete sich grundsätzlich ab, dass in Zukunft ein Miteinander von automatischer und intellektueller Erschließung erwartet wird, zumal das immer größer werdende Publikationsaufkommen teilweise nur noch automatisch bewältigt werden kann. Es wurden Wege diskutiert, wie sich diese beiden Herangehensweisen sinnvoll ergänzen könnten, z.B. durch eine kluge Auswahl der Anwendungsgebiete und durch das Training von maschinellen Verfahren durch Expert/inn/en der Sacherschließung. Potenzial wurde auch in der Vernetzung, der Crosskonkordanzbildung und in der semantischen Verknüpfung gesehen.

Im Anschluss an den ersten Themenblock fanden die beiden Impulsvorträge statt: Im ersten Impulsvortrag „Erschließung für Discovery-Systeme gestalten" skizzierte Jan Maas Anforderungen an die Sach- und Formalerschließung, die sich aus der aktuellen Verwendung von auf Suchmaschinentechnologie basierender Discoverysysteme ergeben. Eine zentrale Idee des Vortrags war, die Erschließung immer mit Blick auf das aktuelle Rechercheparadigma hin zu gestalten, was auf Grund der zahlreichen Technologiewechsel (Zettelkatalog – OPAC – Discoverysystem – ggf. Knowledge Graph) stark erschwert sei. Maas unterteilte die Anforderungen in die drei Bereiche Suche, Relevanzsortierung und Darstellung bzw. Verknüpfung. Speziell bei Sacherschließungsmerkmalen sei eine hohe Abdeckung aller Datensätze eines Discoverysystems wünschenswert bis notwendig. Die Interpretierbarkeit von Feldern durch automatische Verfahren sei nicht immer gegeben, bilde aber eine zentrale Voraussetzung für die Verwendung von Erschließungsmerkmalen in aktuellen und zukünftigen Recherchesystemen.

Maas regte einen konstruktiven Diskurs zwischen Spezialist/inn/en für Sacherschließung und Spezialist/inn/en für Discoverysysteme an, bei dem auch moderne Methoden des Anforderungsmanagements – zum Beispiel Scenario-Based Design, Personas und Storytelling zur Anwendung kommen könnten. Damit könnte eine gemeinsame Diskussionsbasis zwischen Expert/inn/en verschiedener

Bereiche geschaffen werden, denn anhand konkreter Beispiele falle die Kommunikation „über den eigenen Tellerrand hinaus“ leichter.

Im zweiten Lightning Talk stellte Michael Franke-Maier die bisherigen Ergebnisse des Expertenteams „RDA-Anwendungsprofil für die verbale Inhaltserschließung“ (ET RAVI) zur Qualität der Inhaltserschließung vor. Im Auftrag des Standardisierungsausschusses beschäftigt sich dieses Expertenteam seit ca. einem Jahr mit der grundsätzlichen Frage, welche Anforderungen erfüllt sein müssen, damit qualitative Inhaltserschließung gelingen kann. Als Grunddimensionen wurden dabei Transparenz und Verlässlichkeit identifiziert. Transparenz bedeutet hierbei die Offenlegung der Erschließungsmethode, des Erschließungslevels sowie etwaiger Erschließungslücken. Unter Verlässlichkeit versteht das ET RAVI die Konsistenz der Daten, der Regeln und des Retrievals. Dazu kommen drei weitere Dimensionen: Die erste Dimension zu konsistenten Verwendungsregeln zur Ressourcenbeschreibung beinhalte z.B. Anforderungen zur Erschließungstiefe. Franke-Maier stellte die Frage in den Raum, ob eine Flexibilisierung der Erschließungstiefe und die Ermöglichung von enger und weiter Verschlagwortung je nach Bedarfslage, also eine Abkehr von der klassischen Anforderung Präzision, sinnvoll sei? Bei hinreichender Relationierung in den Normdaten sei doch aus einer präzisen inhaltlichen Erschließung eine weite Erschließung skalierbar. Als weitere Anforderung nannte er die Kompatibilität von intellektuellen und maschinellen Verfahren und stellte die Frage, ob es erstrebenswert sei, dass beide Verfahren ähnliche Ergebnisse lieferten. Wäre es nicht viel sinnvoller, die Stärken der beiden Verfahren für ein optimales Retrievalergebnis zu kombinieren? Die zweite Dimension sei eine regelbasierte Produktion von Normdaten als Grundlage für die qualitative Ressourcenbeschreibung. Im Fokus steht hier die Bedeutsamkeit von Wissensorganisationssystemen wie Thesauri und Normdateien mit Anforderungen wie Eindeutigkeit, Disambiguierung, Gebräuchlichkeit und Aktualität sowie die Relationierung innerhalb solcher Systeme und untereinander via Konkordanzen. Die dritte Dimension beschäftige sich schließlich mit der transparenten Auswertung für Retrieval und Anzeige. Die erstellten Daten müssten benutzerfreundlich visualisiert werden, was vor allem aufgrund der Datenheterogenität in so genannten Mega-Indizes eine große Herausforderung sei. Weiterhin müsste das Retrieval neben einer sinnvollen Facettierung auch Möglichkeiten des Navigierens in Normdaten bzw. das thematische Browsen ermöglichen. Er schloss den Kurzvortrag mit der Frage zur Messbarkeit der genannten Anforderungen.

Anschließend folgte die zweite Diskussionsphase mit folgenden Themen:

- Herausforderungen für Erschließungsregeln und Qualitätsanspruch
- Erschließung für Discovery-Systeme gestalten
- Vom Suchen und Finden – die Zukunft nutzerorientierter Recherche

Viele Teilnehmende sprachen sich für klar definierte Qualitätskriterien in der Erschließung aus, die auch unterschiedliche Quellen wie z.B. maschinelle Erschließung berücksichtigen. Im Zusammenspiel von Discoverysystemen und Sacherschließung wurden noch viele Mängel gesehen, die letztendlich durch nachhaltige Entwicklung und intensive Zusammenarbeit von Expert/inn/en beider Gebiete behoben werden müssen. Zukünftige Discoverysysteme müssen sowohl die exakte Suche – bisher eine Stärke klassischer OPACs – als auch die namensgebende explorative Suche gut unterstützen.

Weiterhin wurde gewünscht, die Funktionsweise von Discoverysystemen für Nutzer/inn/en transparenter zu gestalten. Auch innovative Funktionen, die die Sacherschließungsmerkmale besser nutzen, wurden diskutiert, z.B..Ergebnisvisualisierungen oder verbesserte More-Like-This-Funktionen.

Die zahlreichen konstruktiven Redebeiträge und die engagierten Diskussionsteilnehmer/inn/en machten das Knowledge Café zu einer ausgesprochen spannenden Veranstaltung.

Werkstattgespräche maschinelle Erschließung

Der von Christa Schöning-Walter moderierte Workshop begann mit fünf Kurzvorträgen. Vier davon thematisierten die bei der DNB eingesetzten Verfahren und verfolgten Strategien: Sandro Uhlmann erläuterte die maschinelle Beschlagwortung mit GND-Schlagwörtern und Library of Congress Subject Headings (LCSH). Letztere werden derzeit für englischsprachige Hochschulschriften der Reihe O angewendet, wobei für die Named Entities auch hier die GND zum Einsatz kommt. Als Hauptproblem sieht Uhlmann die Disambiguierung an. Beispielsweise wurde eine Dissertation „Empirical essays on the role of stars in collaborative organizations“ fälschlich der Astronomie zugeschlagen und mit „Stars“ (Sterne) beschlagwortet anstatt mit „Celebrities“. Die dahinterstehende Wörterbucharbeit für die maschinelle Beschlagwortung wurde von Jan-Helge Jacobs näher erklärt. Nur bei den Sachbegriffen wird auf der Ebene der kleinsten Zerlegungseinheit (Segment) gematcht, um möglichst viele grammatikalische Varianten zu berücksichtigen. Bei anderen Satzarten (z.B. Personen) werden nur Genitivformen in die Erkennung mit einbezogen. Auch die in GND-Hinweissätzen enthaltenen Informationen werden genutzt, z.B. ergibt „Schriftspracherwerb“ die beiden Schlagwörter „Schriftsprache“ und „Spracherwerb“.

Frank Busse berichtete über die automatische Klassifikation mit Support Vector Machines auf der Averbis Extraction Plattform. Für deutsch- und englischsprachige Netzpublikationen werden sowohl DNB-Sachgruppen als auch DDC-Kurznotationen maschinell vergeben. Drei Prozesse sind zu unterscheiden: das Training, die Routine (d.h. die Verarbeitung des täglichen Zugangs) und die retrospektive Erschließung. Letztere umfasst sowohl die Bearbeitung bisher nicht erschlossener Objekte als auch ggf. eine nochmalige Erschließung mit verbesserten Modellen. Schließlich wurden spannende Einblicke in die technischen Prozesse gegeben. Wie Claudia Grote erläuterte, sind derzeit 39 auf der Averbis-Plattform gespeicherte Verarbeitungsketten produktiv, um 22 Funktionen auszuführen (Vergabe von Sprachcodes, Sachgruppen, Schlagwörter für unterschiedliche Objektarten, Kurznotationen für die einzelnen Sachgruppen). Für jede Verarbeitungskette wird in einem nächtlichen Batch-Verfahren eine Liste mit Identnummern abgearbeitet. Im Durchschnitt werden 7.000 Dokumente behandelt, häufig mit mehreren Verarbeitungen. Dies alles wird bisher von einer selbstentwickelten Verarbeitungssteuerung gelenkt, die mit der Averbis-Plattform kommuniziert. Da bestimmte Schritte jedoch in vielen Verarbeitungsketten vorkommen und deshalb mehrfach ausgeführt werden, strebt die DNB eine Modularisierung mit einer offeneren und flexibleren Architektur an.

Einen interessanten Kontrapunkt setzte der Vortrag von Moritz Fürneisen über die Automatisierung der Sacherschließung beim Leibniz-Informationszentrum Wirtschaft (ZBW). Das erklärte Ziel ist: “So viele Abläufe wie möglich im Erschließungsworkflow automatisieren und dabei die Qualität der von

der ZBW generierten Metadaten erhalten". Der zeitweilig angestrebte Einsatz kommerzieller Lösungen wurde verworfen; seit 2014 wird stattdessen auf eine Eigenentwicklung auf Open-Source-Basis gesetzt. Ein wichtiges Charakteristikum der ZBW-Strategie ist die Kombination mehrerer unterschiedlicher Machine-Learning-Verfahren (derzeit statistische und lexikalische Methoden), um zu verlässlicheren Ergebnissen zu kommen. Künftig soll auch mit neuronalen Netzen und Deep Learning gearbeitet werden. Für die intellektuelle Evaluierung der Ergebnisse wurde ein spezielles Tool entwickelt, das auch die Möglichkeit bietet, fehlende Deskriptoren aus dem Standard-Thesaurus Wirtschaft (STW) zu ergänzen. Außerdem wird an einer automatisierten Qualitätsabschätzung gearbeitet, um in Zukunft maschinell diejenigen Erschließungen zu identifizieren, bei denen eine intellektuelle Nachbearbeitung notwendig ist. Deutlich wurde bei Fürneisens Vortrag, dass die Rahmenbedingungen bei der ZBW etwas günstiger sind als bei der DNB: So stammen die zu erschließenden Publikationen aus klar begrenzten Fachgebieten und oftmals sind Autoren-Keywords als Basis für die Erschließung vorhanden; auch enthält der STW nur wenige Named Entities (primär Geografika).

Im Anschluss an die Kurzvorträge standen alle Referent/inn/en für Detailfragen und Diskussionen zur Verfügung. Dafür waren im Foyer des großen Vortragsraums entsprechende Stationen aufgebaut, an denen man sich in kleinen Gruppen traf.

Ein Skateboard für den Papst – oder Warum es maschinelles Lernen ohne Semantik so schwer hat

Nach den Workshops kamen alle Teilnehmerinnen und Teilnehmer nochmals im Plenum zusammen. Den abschließenden Vortrag hielt Harald Sack vom FIZ Karlsruhe. Er ist Professor am Institut für Angewandte Informatik und Formale Beschreibungsverfahren am Karlsruher Institut für Technologie (KIT) und leitet den Bereich für Information Service Engineering.

Der Vortrag warf ein kritisches Licht auf aktuelle Entwicklungen aus dem Bereich der KI und betrachtete ihre Potenziale, Probleme und Grenzen. Ein besonders wichtiger Aspekt war dabei die Kombination subsymbolischer maschineller Lernverfahren (Deep Learning) mit symbolischen Wissensrepräsentationen (explizite Semantik) zur Einbeziehung von Kontext, Pragmatik und Erfahrungswissen in den Prozess der automatisierten Klassifikation und Erschließung. Insbesondere das Deep Learning werde heute als KI verstanden, wobei in der öffentlichen Wahrnehmung oft aufgrund einzelner, auf eng fokussierte Anwendungsbereiche beschränkter Fähigkeiten von KI-Systemen unzulässige Verallgemeinerungen getroffen würden. So lautete etwa eine Schlagzeile: „Google AI detects breast cancer better than humans". Harald Sack stellte kurz die Geschichte des maschinellen Lernens dar, von den Anfängen der 1940er-Jahre über den „KI-Winter" der 1970er- und 1980er-Jahre bis in das aktuelle Jahrzehnt. Insgesamt stellt diese Entwicklung den Siegeszug des konnektionistischen Paradigmas, also die simulierte Systemordnung mit künstlichen neuronalen Netzen, dar. Wichtige Voraussetzungen für die Weiterentwicklung des Deep Learning, so Sack, war die zunehmende Verfügbarkeit billiger Rechenkapazität mittels Grafikprozessoren und die Verfügbarkeit großer Mengen annotierter Trainingsdaten.

Harald Sack zeigte dem interessierten Publikum verschiedene Anwendungen des Deep Learning für die visuelle Analyse, zum Beispiel die Umsetzung einer Fotografie in den Malstil von Monet oder van Gogh, das Schärfen unscharfer Bilder oder das maschinelle Erkennen und Verschlagworten von Bildinhalten. Ein Ergebnis aus dem letzteren Bereich erklärt auch den Titel des Vortrags: So wurde bei der Anwendung von Bilderkennungssoftware auf mittelalterliche Handschriften ein Bischofsstab in einer Miniatur aus einer Handschrift des 12. Jahrhunderts vom Algorithmus als Skateboard fehlgedeutet. Sack folgerte daraus, dass für eine korrekte Zuordnung zusätzliche semantische Informationen (z.B. der Zeitpunkt der Erfindung von Skateboards) nötig sind, um die Ergebnisse zu verbessern. Das Fazit des Vortrags war, dass Deep Learning unser Leben weiter verändern wird. Out-of-the-box-Modelle seien einfach nutzbar und funktionierten schon recht gut. Entscheidend seien aber die Trainingsdaten, wobei Semantik die Resultate verbessern könne.

Fazit

Die Fachtagung bot eine sehr gelungene Mischung sowohl im Hinblick auf die Formate als auch auf die Inhalte. Sehr anregend war die Abwechslung zwischen Vorträgen einerseits und Workshops bzw. Hands-on Labs andererseits. Auch die Auswahl der Referent/inn/en war spannend, da sie vielfach einen Blick über den Tellerrand des engeren bibliothekarischen Bereichs ermöglichte. Mit besonderer Neugier sahen die Teilnehmenden dabei auf die Firmenvertreter, die freilich – wie es eine Teilnehmerin im Pausengespräch ausdrückte – „auch nur mit Wasser kochen“. Sehr hilfreich für viele Zuhörende waren außerdem die gut verständlichen Erläuterungen zum Thema KI in mehreren Vorträgen, die manchen “Aha-Effekt” ausgelöst haben dürften. Zum Gelingen der Tagung trug aber auch bei, dass die Organisator/inn/en ausreichend Raum für Austausch vorgesehen hatten: Nicht nur im Anschluss an die Vorträge wurde eifrig diskutiert, sondern auch in den Workshops, in den Pausen und beim gemeinsamen Abendessen am ersten Tag. Der DNB sei ein großer Dank für die Ausrichtung der Fachtagung ausgesprochen, die mit einem Schlusswort von Ulrike Junger (DNB) endete.

Michael Franke-Maier, Universitätsbibliothek der Freien Universität Berlin
Cyrus Beck, Zentralbibliothek Zürich
Anna Kasprzik, ZBW – Leibniz-Informationszentrum Wirtschaft, Hamburg
Jan Frederik Maas, Staats- und Universitätsbibliothek Hamburg
Sarah Pielmeier, Universitäts- und Landesbibliothek Münster
Heidrun Wiesenmüller, Hochschule der Medien Stuttgart

Zitierfähiger Link (DOI): https://doi.org/10.5282/o-bib/5565