Automatisierte semantische Anreicherung von historischen Texten

Erkennung und Verknüpfung von Entitäten mit Wikidata und Wikipedia

Kai Labusch, Sophie Schneider, Clemens Neudecker

Abstract

Die Zugänglichkeit zunehmend digital verfügbarer historischer Bestände der Staatsbibliothek zu Berlin (SBB) soll auf Grundlage des Inhalts der Werke zukünftig weiter ausgebaut werden. Im Referat Data Science wurden entsprechende Lösungen erarbeitet, mit deren Hilfe, basierend auf der Erkennung und Verlinkung von in den Texten genannten Entitäten, automatisch semantische Anreicherungen generiert werden können. Der Beitrag beschreibt, wie unter Zuhilfenahme von Technologien aus dem Bereich der Künstlichen Intelligenz bzw. des Maschinellen Lernens, sowie Wikidata und Wikipedia als Wissensdatenbanken, Entitäten identifiziert, klassifiziert und verknüpft werden. Wir stellen eine Fallstudie zum Topic Modeling mit Named Entities vor und zeigen Zukunftsperspektiven für die Weiterentwicklung auf.

The accessibility of the increasingly digitally available historical collections of the Berlin State Library (SBB) can be expanded in the future based on textual contents. Within the Data Science department appropriate solutions for the recognition and linking of Named Entities and thus automatically generated semantic enrichments have been developed. The article describes how entities are being identified, classified and linked with the help of technologies from the field of artificial intelligence and machine learning, employing Wikidata and Wikipedia as knowledge bases. We present a case study on Topic Modeling with Named Entities and discuss future perspectives for further development.

Einleitung

Die Staatsbibliothek zu Berlin (SBB) als Teil der Stiftung Preußischer Kulturbesitz (SPK) treibt seit über 15 Jahren die Digitalisierung ihrer Bestände voran, um diese online den Nutzenden zugänglich zu machen. Die Digitalisierung erfolgt größtenteils durch das hauseigene Digitalisierungszentrum, das seit 2007 existiert. Hier werden kontinuierlich Bücher, Zeitungen, Handschriften und dergleichen gescannt. Aktuell ist ein wachsendes Datenkorpus von mehr als 7 Peta-Bytes vorhanden, das mehr als 200.000 digitalisierte Werke umfasst, wobei bislang nur für etwa 5 Millionen Seiten Volltexte mittels optischer Zeichenerkennung (OCR) generiert wurden. Die OCR-Volltexte werden im ALTO-Format[1] bereitgestellt und unter anderem für die Volltextindizierung verwendet. Nutzende können Digitalisate über das Online-Portal der Digitalisierten Sammlungen[2] der SBB abrufen und dort neben den Metadaten auch die Volltexte durchsuchen.

Im Rahmen des Qurator Projektes[3], das durch das Bundesministerium für Bildung und Forschung (BMBF) gefördert wurde, ist an der SBB unter anderem Software für die Layoutanalyse komplexer mehrspaltiger Dokumente, wie z.B. historische Zeitungen, entstanden. Weitere Resultate des Qurator Projektes sind Methoden zur semantischen Anreicherung von Volltexten, d.h. die Erkennung und Klassifikation von Eigennamen bzw. benannten Entitäten im OCR-Volltext (Named Entity Recognition, NER) sowie für die automatisierte Verknüpfung der erkannten Entitäten (Entity Linking, EL) mit einer Wissensdatenbank, in diesem Fall Wikidata.

Durch die Etablierung einer in-house OCR-Pipeline auf Basis von OCR-D[4] soll in Zukunft für alle digitalisierten Drucke der SBB ein entsprechender OCR-Volltext verfügbar gemacht werden. Volltextbestände digitalisierter historischer Werke können durch weitergehende semantische Anreicherungen automatisiert tiefer erschlossen werden. Die semantische Anreicherung ermöglicht dann die Schaffung neuer Zugänge zu den Beständen für die Nutzenden. Die Entwicklungsarbeiten hierzu werden im aktuell laufenden Projekt Mensch.Maschine.Kultur (MMK)[5] fortgeführt, das durch die Beauftragte des Bundes für Kultur und Medien (BKM) für drei Jahre gefördert wird. Im Fokus der Arbeiten zur semantischen Anreicherung steht das Volltextkorpus der Digitalisierten Sammlungen.

Dieser Artikel soll einen Überblick über den aktuellen Stand der Arbeiten bieten. Zunächst möchten wir motivieren, warum wir diese Ansätze verfolgen. Als Nächstes

1 https://www.loc.gov/standards/alto/ [22.04.2024].
2 https://digital.staatsbibliothek-berlin.de/ [22.04.2024].
3 https://qurator.ai/ [22.04.2024].
4 https://ocr-d.de/ [22.04.2024].
5 https://mmk.sbb.berlin/ [22.04.2024].

beschreiben wir den Weg vom analogen Dokument zum digitalen Volltext. Dann erläutern wir, wie die Textpassagen, die Entitäten bezeichnen, in dem Volltext gefunden werden. Im Weiteren beschreiben wir detailliert, wie die automatisierte Verknüpfung der Entitätsbezeichner mit einer Wissensdatenbank wie Wikidata oder der Gemeinsamen Normdatei (GND) funktioniert. Danach stellen wir als Beispiel für die Verwendung der resultierenden semantisch angereicherten Texte eine Fallstudie zum Topic-Modeling basierend auf Entitäten vor. Zum Schluss geben wir Hinweise zu Daten- und Quellcode-Ressourcen sowie einen Ausblick in die Zukunft der semantischen Anreicherung in den Digitalisierten Sammlungen der SBB.

Motivation: Neue Zugangsperspektiven

Warum sind wir an der Erkennung, Disambiguierung und Verknüpfung von Entitäten interessiert? Und was ist der Nutzen einer solchen semantischen Anreicherung, wenn schon Volltexte vorliegen?

Im Zentrum unserer Überlegungen stehen die Nutzenden der SBB. Sie sollen in die Lage versetzt werden, strukturierte semantische Informationen, wie sie in der GND oder in Wikidata vorliegen, zu verwenden, um gezielten Zugriff auf für sie relevante Textpassagen in den OCR-Volltexten zu erhalten. Beispiel sei hier eine Nutzerin der SBB, die auf der Suche nach Texten in den Digitalisierten Sammlungen ist, die Musiker erwähnen, deren Geburtsort Rotterdam ist. Schon jetzt wäre es möglich, über eine Recherche die Namen entsprechender Musiker zu identifizieren sowie über die Volltextsuche der Digitalisierten Sammlungen anschließend danach zu suchen. Allerdings zeichnet sich dabei vermutlich schnell folgendes Problem ab: die einfache Suche der Nutzerin ergibt zwar möglicherweise für sie relevante Treffer, aber auch viele (a) Treffer, welche den entsprechenden Suchbegriff enthalten, aber gar keine Personen sind, weil sie einer anderen oder von uns bislang nicht definierten Klasse von Entitäten angehören oder (b) Treffer, die zwar Personen sind und z.B. den gleichen (Nach-)Namen tragen, aber nicht mit der gesuchten Person übereinstimmen (andere Identität). Named Entity Recognition (a) und Entity Linking (b) Informationen können hier Abhilfe schaffen, indem die Ergebnismenge erheblich auf die wahrscheinlichste gesuchte Entität eingeschränkt werden kann.

Des Weiteren sind durch die Einbindung von NER/EL-Ergebnissen auch komplexere Anfragen an die Digitalisierten Sammlungen denkbar, bei denen die genaue Identität der gesuchten Person(en) noch nicht vorab bekannt sein muss. So könnte die Nutzerin zu ihrer Fragestellung eine passende SPARQL-Abfrage (SPARQL Protocol And RDF Query Language) formulieren, die über eine Wikidata-Abfrage eine Liste von Personenentitäten liefert, die einerseits Musiker sind oder waren und andererseits in Rotterdam geboren wurden. Diese Liste wiederum kann dann automatisiert gegen alle verknüpften Entitäten in den Volltexten der Digitalisierten Sammlungen abgeglichen werden und somit können der Nutzerin Textpassagen mit den gewünschten Eigenschaften angezeigt werden.

Vom digitalen Bild zum Volltext

Vorbedingung für die semantische Anreicherung ist ein vorliegender Volltext von guter Qualität. Dieser Volltext entsteht als Ergebnis einer Reihe von Verarbeitungsschritten. Zunächst werden im Digitalisierungszentrum Dokumente gescannt, so dass die einzelnen Seiten als digitale Bilder vorliegen. Danach werden die erzeugten Bilder einer Layout-Analyse (oder auch Segmentierung) unterzogen. Hierbei wird das Layout der Seite analysiert und dessen Struktur mittels Methoden des maschinellen Lernens (ML) erkannt, d.h. der logische Aufbau der Seite – Überschriften, Textblöcke, Separatoren, Bilder, Abbildungen, Tabellen, Marginalien, Initialen usw. – wird bestimmt. Dies geschieht mit einer speziell für historische Dokumente an der Staatsbibliothek entwickelten Software, welche auf pixelbasierter Segmentierung beruht[6]. Auf Basis der ermittelten Seitenstruktur wird der Text der einzelnen Strukturelemente mittels OCR erkannt. In einem letzten Schritt wird die Lesereihenfolge des Dokuments[7], wiederum unter Verwendung von ML-Werkzeugen, automatisiert bestimmt.

Am Ende der Verarbeitungskette steht ein digitalisiertes Dokument mit erkannter Seitenstruktur und zugehörigem Volltext sowie definierter Lesereihenfolge.

Erkennung von Entitäten

Die Klassen von Entitäten, an denen wir im Folgenden interessiert sind, sind Personen, Orte und Organisationen (Körperschaften). Zunächst müssen also im Volltext Bezeichner von Personen, Orten und Organisationen identifiziert werden. Dieser Schritt wird als Entitätenerkennung oder Named Entity Recognition (NER) bezeichnet.

NER findet im Text die Passagen, die sich auf eine Person, einen Ort oder eine Organisation beziehen. An der SBB wird für diesen Zweck ein auf BERT (Bidirectional Encoder

6 Rezanezhad, Vahid / Baierer, Konstantin / Gerber, Mike / Labusch, Kai / Neudecker, Clemens: Document Layout Analysis with Deep Learning and Heuristics, in: Proceedings of the 7th International Workshop on Historical Document Imaging and Processing (HIP,23) 2023, S. 73–78. https://doi.org/10.1145/3604951.3605513. Quellcode: https://github.com/qurator-spk/eynollah [22.04.2024].

7 Vgl. Clausner, Christian / Pletschacher, Stefan / Antonacopoulos, Apostolos: The significance of reading order in document recognition and its evaluation, in: 12th International Conference on Document Analysis and Recognition 2013. S. 688-692. http://dx.doi.org/10.1109%2FICDAR.2013.141

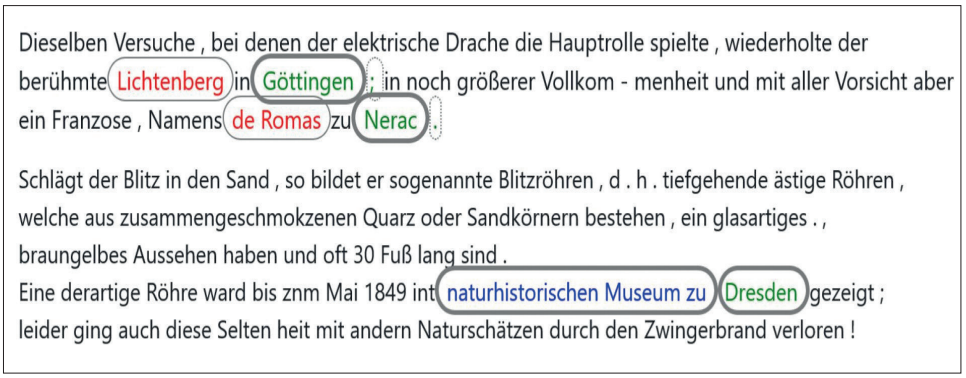

Abbildung 1: OCR-Volltextausschnitt aus „Das Buch der wunderbaren Erfindungen" von Thomas Louis (1860). Ergebnis der NER: Personenbezeichner(rot), Ortsbezeichner (grün), Organisationsbezeichner (blau). Die Stärke der Umrandung visualisiert die Konfidenz des Entity-Linking – je stärker die Umrandung desto höher ist die Konfidenz der Verknüpfung.

- Adelheid von Lichtenberg I. († 1312), deutsche Äbtissin
- Adelheid von Lichtenberg II. († ca. 1413), deutsche Äbtissin
- Agnes von Lichtenberg I. († 1336), deutsche Äbtissin
- Agnes von Lichtenberg II. († 1436), deutsche Äbtissin
- Alexander von Lichtenberg (1880–1949), ungarischer Urologe
- Anna von Lichtenberg (1442–1474), Erbtochter der Herrschaft Lichtenberg
- Bernd Lichtenberg (* 1966), deutscher Drehbuchautor
- Bernhard Lichtenberg (1875–1943), deutscher Seliger und Gerechter unter den Völkern
- Betz von Lichtenberg († 1480), Großbailli
- Byron Kurt Lichtenberg (* 1948), US-amerikanischer Astronaut
- Carl Lichtenberg (1816–1883), deutscher Politiker
- Claudia Lichtenberg (* 1985), deutsche Radrennfahrerin
- Dieter Lichtenberg (* 1949), deutscher Kanute
- Eleonora Jakowlewna Lichtenberg (* 1925), sowjetisch-russische Architektin
- Elisabeth von Lichtenberg († 1320), deutsche Äbtissin
- Erik Lichtenberg, US-amerikanischer Agrarökonom
- Ernst Lichtenberg (* 1939), deutscher Politiker (CDU)
- Friedrich Lichtenberg (1801–1871), deutscher Politiker, MdL Hessen
- Friedrich David Lichtenberg (1774–1847), deutscher Apotheker
- Friedrich August von Lichtenberg (1755–1819), deutscher Politiker
- Georg Lichtenberg (Politiker) (1852–1908), deutscher Politiker, Bürgermeister von Linden
- Georg Lichtenberg (Landrat) (1886–1973), deutscher Verwaltungsjurist, Landrat von Neustadt
- Georg Christoph Lichtenberg (1742–1799), deutscher Physiker und Aphoristiker
- Gustav Wilhelm Lichtenberg (1811–1879), deutscher Jurist und Politiker, MdL Hessen
- Hagen Lichtenberg (* 1943), deutscher Jurist
- Hans-Jürgen Lichtenberg (* 1940), deutscher Politiker
- Hermann II. Hummel von Lichtenberg († 1335), Bischof von Würzburg und Herzog in Franken
- Jacqueline Lichtenberg (* 1942), US-amerikanische Schriftstellerin
- Jakob von Lichtenberg (auch *Jakob im Bart*; 1416–1480), Vogt von Straßburg
- Jessica Lichtenberg (* 1981), deutsche Voltigiererin
- Johann Conrad Lichtenberg (1689–1751), deutscher Theologe, Generalsuperintendent, Baumeister und Librettist
- Joseph D. Lichtenberg (1925–2021), US-amerikanischer Psychiater und Psychoanalytiker
- Lorenz von Lichtenberg (Bischof, † 1279) (*Lorenz von Leistenberg*; † 1279), Bischof von Metz
- Lorenz von Lichtenberg (Bischof, † 1446) (*Lorenz II. von Lichtenberg*; † 1446), Bischof von Lavant und von Gurk
- Ludwig Christian Lichtenberg (1737–1812), deutscher Beamter und Naturforscher
- Ludwig von Lichtenberg (1784–1845), deutscher Politiker
- Mechtelt van Lichtenberg (um 1520–1598), niederländische Malerin
- Meza von Lichtenberg († 1263), deutsche Äbtissin
- Mike Leon Lichtenberg (* 2001), deutscher Schauspieler und Musiker
- Otto Friedrich von Lichtenberg († 1838), deutscher Politiker und Gutsbesitzer auf Scherneck
- Paul Lichtenberg (1911–1995), deutscher Bankmanager
- Reinhold von Lichtenberg (1865–1927), österreichischer Kunsthistoriker und Publizist
- Ruby M. Lichtenberg (* 2005), deutsche Schauspielerin
- Simon Lichtenberg (* 1997), deutscher Snookerspieler
- Uwe Lichtenberg (1934–2011), deutscher Politiker, Oberbürgermeister von Fürth
- Werner Lichtenberg (* 1953), deutscher Verwaltungsjurist
- Wilhelm Lichtenberg (1892–1960), Schweizer Dramatiker

Abbildung 2: Aus Wikipedia – Liste der Namensträger „Lichtenberg"

Representations from Transformers)[8] basierendes NER-Modell verwendet. Hierzu wurde ein BERT-Modell entwickelt, das für historische und mit OCR-Fehlern behaftete Texte in deutscher Sprache optimiert ist[9]. Das Modell wurde dazu auf den bereits vorhandenen Volltexten der Digitalisierten Sammlungen vortrainiert, um es an das spezielle Textmaterial mit seinen historischen Schreibvarianten sowie noch vorhandene OCR-Fehler zu adaptieren. Die Feinabstimmung auf die NER-Aufgabe erfolgte dann für deutschsprachige Texte mit zeitgenössischen und historischen deutschsprachigen Ground Truth[10] Daten. Für französische, englische und niederländische Texte wurde eine Kombination aus Ground Truth Daten für diese Sprachen und deutschsprachiger Ground Truth verwendet. Im Endresultat entstanden dabei NER-Modelle für OCR-Volltexte historischer Dokumente in deutscher, französischer und englischer Sprache[11].

Nachdem die NER erfolgt ist, liegen digitalisierte Dokumente mit Seitenstruktur, Volltexten, Lesereihenfolge sowie zusätzlich identifizierten Entitätsbezeichnern vor. Zusätzlich zum Text gibt es nun also markierte Textpassagen, die sich jeweils auf eine Person, einen Ort oder eine Orga-

8 Devlin, Jacob / Chang, Ming-Wei / Lee, Kenton / Toutanova, Kristina: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 2019, S. 4171-4186. https://doi.org/10.18653/v1/N19-1423.

9 Labusch, Kai / Neudecker, Clemens / Zellhöfer, David: BERT for Named Entity Recognition in Contemporary and Historical German, in: Proceedings of the 15th conference on natural language processing 2019. S. 8-11. https://konvens.org/proceedings/2019/papers/KONVENS2019_paper_4.pdf.

10 Als Ground Truth bezeichnet man durch manuelle Annotation erstellte korrekte Daten, hier also Volltexte, in denen sämtliche Vorkommen von Personen, Orten und Organisationen nach Annotationsrichtlinien erfasst wurden.

11 Vgl. Labusch, Kai / Neudecker, Clemens: Named Entity Disambiguation and Linking Historic Newspaper OCR with BERT, in: CLEF 2020. https://ceur-ws.org/Vol-2696/paper_163.pdf. Menzel, Sina / Schnaitter, Hannes / Zinck, Josefine / Petras, Vivien / Neudecker, Clemens / Labusch, Kai / Leitner, Elena / Rehm, Georg: Named Entity Linking mit Wikidata und GND – Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten, in: Franke-Maier, Michael / Kasprzik, Anna / Ledl, Andreas / Schürmann, Hans (Hrsg.): Qualität in der Inhaltserschließung, Berlin, Boston: De Gruyter Saur 2021. S. 229-258. https://doi.org/10.1515/9783110691597-012.

nisation beziehen. In einem weiteren Schritt erfolgt nun die Disambiguierung und Entitätenverknüpfung.

Verknüpfung von Entitäten

Entitätenverknüpfung oder Entity Linking (EL) entspricht prinzipiell dem Nachschlagen in einer Wissensdatenbank wie der GND, Wikidata oder Wikipedia. Hierzu betrachten wir das Beispiel in Abbildung 1. Diese zeigt einen Ausschnitt aus einem Volltext. Die farbigen Markierungen des Textes entsprechen der Ausgabe des BERT-basierten NER-Systems. Hier wurden zwei Personenbezeichner – „Lichtenberg" und „de Romas" – drei Ortsbezeichner – „Göttingen", „Nerac" und „Dresden" – sowie ein Organisationsbezeichner „naturhistorisches Museum" – identifiziert. Beim Entitätendisambiguierungs- und Verknüpfungsschritt geht es nun darum, z.B. für den Personenbezeichner „Lichtenberg" zu bestimmen, ob hiermit eine Person gemeint ist, die einem konkreten Eintrag in einer Wissensbasis wie der GND oder Wikidata entspricht. Die naive Suche für Personen mit Nachnamen „Lichtenberg" liefert z.B. in der Wikipedia eine Liste mit ca. 50 Einträgen aus. Abbildung 2 zeigt einen Ausschnitt dieser Suchresultate. Ziel der Entitätenverknüpfung und -disambiguierung ist es, möglichst wenige Kandidaten für die Verknüpfung mit einer ausreichend großen Wahrscheinlichkeit zu ermitteln. Was einer ausreichend großen Wahrscheinlichkeit entspricht, wird mithilfe eines Schwellwertes definiert. Alle Einträge der Wissensbasis mit einer Verknüpfungswahrscheinlichkeit unterhalb dieses Wertes werden nicht gelistet. Die resultierende Liste an Kandidaten wird noch nach Konfidenzen sortiert, d.h. sie beginnt mit dem wahrscheinlichsten Kandidaten. Wir möchten also in diesem Fall eine Liste mit lediglich zwei Einträgen erhalten, beispielsweise wie in Abbildung 3, aus der hervorgeht, dass mit der höchsten Wahrscheinlichkeit hier „Georg Christoph Lichtenberg" gemeint ist, mit einer geringeren Wahrscheinlichkeit „Ludwig Christian Lichtenberg" und dass alle anderen „Lichtenbergs" aus der Wissensbasis eine so geringe Verknüpfungswahrscheinlichkeit besitzen, dass sie gar nicht erst mit aufgeführt werden. Natürlich könnte für manche Text-, Orts- oder Organisationsbezeichner das Ergebnis auch sein, dass diese Bezeichner keiner der Entitäten in der Wissensbasis mit ausreichender Wahrscheinlichkeit zugeordnet werden können.

Verknüpfung von Entitäten – Wissensbasis

Bevor Entity Linking geschehen kann, ist es zunächst notwendig, eine Wissensbasis zu definieren. Die Wissensbasis enthält alle Entitäten, die im Rahmen des Entity Linking adressiert werden können. In unserem Fall ist ein elementarer Schritt der Verknüpfung der paarweise Vergleich von Textstellen. Hierbei werden Textpassagen des Zieltextes, in dem die Verknüpfungen vorgenommen werden sollen, mit Textpassagen aus der Wikipedia verglichen, welche

Wikipedia	Wikidata	Confidence
Georg_Christoph_Lichtenberg	Q57554	0.50
Ludwig_Christian_Lichtenberg	Q1874282	0.18

Abbildung 3: Ergebnisliste des Entity-Linking. Es gibt zwei Personenentitäten deren Verknüpfungswahrscheinlichkeiten oberhalb des gewählten Schwellwertes von 0.1 liegen.

die möglichen Verknüpfungskandidaten erwähnen. Somit verwendet das EL-System Informationen, die sowohl aus Wikidata als auch aus Wikipedia stammen. Mittels Wikidata wird die Gesamtheit aller adressierbaren Entitäten ermittelt und Wikipedia liefert zu jeder der adressierbaren Entitäten Textvergleichsmaterial.

Insgesamt nutzen wir einen Satz aus 14 SPARQL-Abfragen, um die Menge der adressierbaren Entitäten in den Klassen Person, Ort und Organisation aus Wikidata zu ermitteln. Wir betrachten dabei alle Unterklassen der Konzepte, die in Abbildung 4 gezeigt werden.

Person	Ort	Organisation
Person(Q215627) Subjekt(Q830077) Fiktiver Charakter (Q95074) Fiktive Person (Q97498056)	Räumliche Entität(Q58416391) Geografische Entität (Q27096213) Fiktiver Ort (Q3895768)	Gruppe von Menschen (Q16334295) Institution (Q178706) Organ(Q895526) Verband (Q15911314) Gewerbebetrieb (Q4830453) Bewaffnete Organisation (Q17149090) Fiktive Organisation (Q14623646)

Abbildung 4: Um die adressierbaren Entitäten der Wissensdatenbank zu ermitteln werden alle Instanzen (P279 Subclass of) von insgesamt 14 Wikidata-Klassen betrachtet.*

```
SELECT ?organisation ?label ?sitelink ?gndid
WHERE
{
  # instance of any subclass of  fictional organisation
  ?organisation wdt:P31/wdt:P279* wd:Q14623646;rdfs:label ?label.
  FILTER(LANG(?label) = "fr").
  ?sitelink schema:about ?organisation.
  FILTER (CONTAINS(str(?sitelink), "fr.wikipedia.org")).
  OPTIONAL {
    ?organisation wdt:P227 ?gndid
  }
}
```

Abbildung 5: Wikidata SPARQL-Abfrage für alle Instanzen der Klasse „fiktionale Organisation"

Abbildung 5 zeigt beispielhaft eine der SPARQL-Abfragen, die genutzt werden, um aus Wikidata die Gesamtheit der adressierbaren Entitäten für die Wissensbasis zu extrahieren. Die abgebildete Beispielabfrage liefert alle Instanzen der Klasse „fiktionale Organisation". Falls vor-

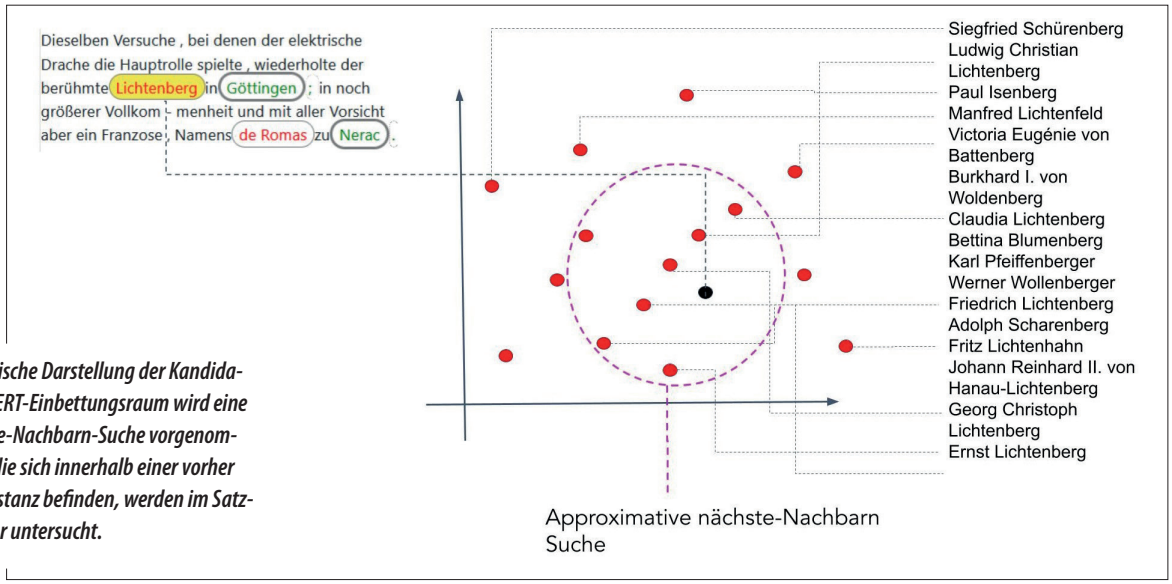

Abbildung 6: Schematische Darstellung der Kandidatenbestimmung. Im BERT-Einbettungsraum wird eine approximative Nächste-Nachbarn-Suche vorgenommen. Die Suchtreffer, die sich innerhalb einer vorher gewählten Maximaldistanz befinden, werden im Satzvergleichsschritt weiter untersucht.

handen, wird für jede Entität aus Wikidata zudem die korrespondierende GND-ID ausgelesen, so dass (sofern diese hinterlegt ist) ebenfalls eine Verknüpfung zur GND erfolgen kann. Es werden hierbei nur diejenigen Entitäten berücksichtigt, für die ein entsprechender Eintrag in der deutschen, französischen oder englischen Wikipedia vorhanden ist. Diese Einschränkung ist notwendig, da in einem zweiten Schritt für jede der Entitäten in der Wissensdatenbank diejenigen Sätze aus der Wikipedia ermittelt werden, die eine Verlinkung zum Wikipedia-Eintrag der Entität enthalten.

Verknüpfung von Entitäten – Verarbeitungsschritte

Die eigentliche Disambiguierung und Verknüpfung der Entitäten erfolgt in drei Schritten. Im ersten Schritt, der Kandidatenbestimmung, wird eine Menge an möglichen Verknüpfungskandidaten aus der Wissensbasis ermittelt. Im zweiten Schritt, dem Satzvergleich, werden eine Reihe von Satzpaaren aus dem Volltext und aus Wikipedia pro Verknüpfungskandidaten verglichen. Im dritten und letzten Schritt, der Bewertung, werden statistische Merkmale der Satzvergleichsresultate pro Verknüpfungskandidat betrachtet, um eine finale Entscheidung über die Verknüpfungswahrscheinlichkeit zu treffen. Im Folgenden erläutern wir diese drei Schritte im Detail.

Verknüpfung von Entitäten – Kandidatenbestimmung

Abbildung 6 zeigt eine schematische Darstellung der Kandidatenbestimmung. Die Kandidatenbestimmung basiert auf einer Nächste-Nachbarn-Suche in einem BERT-Einbettungsraum[12]. Für die Einzelkomponenten der Bezeichner aller adressierbaren Entitäten der Wissensbasis wurden zuvor BERT-Einbettungsvektoren berechnet und in einem Vektorindex[13] gespeichert. Dieser Vektorindex ermöglicht es, ressourceneffizient die nächsten Nachbarn eines gegebenen Einbettungsvektors aus der Gesamtmenge aller Entitäten der Wissensbasis zu ermitteln.

Betrachten wir wieder das Beispiel in Abbildung 6, so wird für die gegebene Textpassage "Lichtenberg" zunächst der korrespondierende BERT-Einbettungsvektor berechnet. Aus den nächsten Nachbarn dieses Einbettungsvektors im vorberechneten Vektorindex wird eine Kandidatenliste ermittelt die mehrstufig nach heuristischen Merkmalen, wie z.B. der a-priori Wahrscheinlichkeit der zugehörigen Entitäten, sortiert wird. Die a-priori-Wahrscheinlichkeit einer Entität ist die Anzahl der Verlinkungen auf die jeweilige Entität in der gesamten Wikipedia, geteilt durch die Gesamtanzahl aller Verlinkungen in der Wikipedia. Die vordersten 20 Einträge dieser sortierten Liste sind dann die Verknüpfungskandidaten, die im Satzvergleichs- und Bewertungsschritt weiter untersucht werden. Mögliche Verknüpfungskandidaten für das Beispiel "Lichtenberg" werden in Abbildung 6 gezeigt.

Verknüpfung von Entitäten – Satzvergleiche

Im Satzvergleichsschritt werden pro zuvor bestimmten Kandidaten bis zu 50 Satzpaare gebildet. Die Satzpaare bestehen jeweils aus einem Satz des Zieltextes, also z.B. dem Satz aus Abbildung 1, der den zu verknüpfenden Bezeichner enthält, z.B. "Lichtenberg", und jeweils einem Satz aus der Wikipedia, der einen Link auf den zu evaluierenden Verknüpfungskandidaten enthält. Abbildung 7 zeigt dies exemplarisch für das "Lichtenberg" Beispiel und eine Reihe von Verknüpfungskandidaten sowie Beispiel-

12 Wiedemann, Gregor / Remus, Steffen / Chawla, Avi / Biemann, Chris: Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. arXiv preprint arXiv:1909.10430, 2019. https://arxiv.org/abs/1909.10430.

13 https://github.com/spotify/annoy [22.4.2022].

Abbildung 7: Schematische Darstellung des Satzvergleichs. Ein zuvor auf diese Aufgabe trainiertes BERT-Modell vergleicht Sätze des Textes, dessen Entitäten verknüpft werden sollen, mit Sätzen aus der Wikipedia, die von Wikipedia-Autoren mit möglichen Kandidaten verknüpft wurden.

sätze aus der Wikipedia, die diese Kandidaten erwähnen. Für jedes der Satzpaare bestimmt dann ein weiteres, angepasstes BERT-Modell, das Evaluierungsmodell, die Wahrscheinlichkeit dafür, dass beide Sätze die identische Entität referenzieren. Das Evaluierungsmodell wurde zuvor für diese Aufgabe trainiert, indem Wikipedia-Satzpaare präsentiert wurden, die entweder Verknüpfungen zur gleichen Entität enthalten oder Verknüpfungen zu unterschiedlichen Entitäten[14].

Verknüpfung von Entitäten – Bewertung

In den vorherigen Schritten, der Kandidatenbestimmung und dem Satzvergleich, wurden für eine fragliche Textstelle bis zu 20 Verknüpfungskandidaten ermittelt und für jeden dieser Verknüpfungskandidaten bis zu 50 Satzvergleiche vorgenommen. Hieraus ergibt sich eine Menge von bis zu 50 Übereinstimmungswahrscheinlichkeiten die durch das Evaluierungsmodell pro Kandidaten ermittelt wurden. Im Bewertungsschritt wird aus statistischen Kennzahlen dieser Menge von Übereinstimmungswahrscheinlichkeiten, wie z.B. Minimum, Maximum, Standardabweichung, Mittelwert, verschiedenen Quantilen, Rangstatistik und weiteren, ein Merkmalsvektor gebildet, auf dessen Grundlage ein zuvor trainierter statistischer Klassifikator, ein Random-Forest-Modell[15], eine Gesamtübereinstimmungswahrscheinlichkeit pro Verknüpfungskandidaten berechnet.

Alle Verknüpfungskandidaten werden absteigend gemäß dieser Gesamtübereinstimmungswahrscheinlichkeit sortiert. Kandidaten, deren Wahrscheinlichkeit unterhalb eines Schwellwertes liegt, werden aus der Liste entfernt. Die abgeschnittene und absteigend nach Übereinstimmungswahrscheinlichkeiten sortierte Kandidatenliste ist schließlich das Endergebnis des Entity Linking.

NER + EL auf den Digitalisierten Sammlungen

Mittels des vorgestellten NER-Verfahrens wurden für ca. 5 Millionen Seiten Volltext aus den Digitalisierten Sammlungen der SBB[16] die Personen-, Orts- und Organisationsbezeichner ermittelt und dann mit dem beschriebenen EL-Verfahren mit der Wissensbasis Wikidata verknüpft. Dabei wurden ca. 65 Millionen Verknüpfungen vorgenommen. Ziel weiterer Arbeiten ist es nun, die um die Entitäten und deren Verknüpfungen angereicherten Volltexte wieder in das Web-Portal der Digitalisierten Sammlungen der SBB zu integrieren und für Suche und Präsentation nutzbar zu machen. Im Zentrum der Überlegungen steht der Gedanke, sowohl das direkte Auffinden von Informationen zu bestimmten Entitäten, als auch die weniger gezielte, entitätengeleitete Erkundung des Bestandes zu optimieren. Umsetzbar wäre dies beispielsweise über entsprechende Filter-/Sortierfunktionen für Ergebnislisten, ähnlichkeitsbasierte Vorschlagsysteme (Recommender Systems) oder eine entitätenzentrierte Sammlungs(re) organisation. Über eine weiterführende Verknüpfung der identifizierten Entitäten innerhalb und zwischen Texten der Digitalisierten Sammlungen sowie auch mit externen Inhalten können über bekannte Sammlungsinhalte neuartige Informationen entdeckt und Querbeziehungen hergestellt werden. In vergleichbarer Weise kann auch die Visualisierung von Entitäten und deren Relationen mittels der Kontextualisierung innerhalb bislang unerforschter (z.B. räumlicher, zeitlicher) Zusammenhänge zu neuen Forschungsfragen führen.

14 Vgl. Labusch, Kai / Neudecker, Clemens: Entity Linking in Multilingual Newspapers and Classical Commentaries with BERT, in: CLEF 2022. S. 1079-1089. https://ceur-ws.org/Vol-3180/paper-85.pdf.

15 Breiman, Leo: Random forests, in: Machine learning 45 (2001). S. 5-32. https://link.springer.com/content/pdf/10.1023/a:1010933404324.pdf [22.04.2024].

16 https://digital.staatsbibliothek-berlin.de/ [22.04.2024].

Das Forschungsfeld der Digital Humanities eröffnet viele potenzielle Anwendungsfelder wie beispielsweise die Historische Netzwerkanalyse (HNA, s.a. SoNAR (IDH)[17]) oder die Untersuchung der Bestände der Digitalisierten Sammlungen mit Methoden der Digital Humanities wie Topic Modeling (welche Entitäten treten in gemeinsamen Themen auf) oder computergestützten Kontext-/Sentimentanalysen (wie wird über eine Entität geschrieben). Im Folgenden möchten wir dazu beispielhaft beschreiben, wie EL der Volltexte genutzt werden kann, um ein ML-getriebenes automatisches Topic Modeling für die deutschsprachigen Werke der Digitalisierten Sammlungen zu erhalten.

Fallstudie – Topic Modeling der Digitalisierten Sammlungen

Ziel von Topic Modeling ist es, innerhalb einer gegebenen Textsammlung, wie z.B. den Digitalisierten Sammlungen der SBB, automatisiert thematische Schwerpunkte zu identifizieren. Das Ergebnis von Topic Modeling Verfahren ist dabei nicht nur abhängig vom zugrunde liegenden Textmaterial, sondern auch vom verwendeten Algorithmus und von benutzerdefinierten Eingabeparametern, wie z.B. der Anzahl der Themengebiete, die zu bestimmen sind. Datengetriebenes automatisiertes Topic Modeling wird hier nicht als Ersatz für die intellektuelle Erschließung des Bestandes betrachtet, sondern als Werkzeug für die Untersuchung und Visualisierung von Eigenschaften großer Textbestände. Lassen sich die Werke des vorliegenden Textkorpus über die Personen, Orte oder Organisationen, die in ihnen erwähnt werden, thematisch gruppieren? Wenn eine solche Gruppierung in einer vorgegebenen Granularität für eine vorgegebene Entitätsklasse bzw. Untermengen von Entitäten vorgenommen wurde, welche inhaltlichen Schwerpunkte lassen sich dann identifizieren und welche Entitäten stehen innerhalb einer Gruppe (Cluster) in einer engeren Beziehung zueinander? Ein etabliertes objektives Maß für die Qualität des Topic Modeling Ergebnis existiert dabei nicht, da das Ergebnis immer im Kontext der jeweiligen Fragestellung bewertet werden muss. Die Nutzenden sollen in die Lage versetzt werden, im Sinne eines "Distant Reading"[18] auf explorative Art und Weise einen Überblick über die Sammlung zu erhalten, der über die Betrachtung von wenigen Einzelwerken hinausgeht.

Die Grundlage von Topic Modeling ist die Abbildung des Inhalts der Werke einer Sammlung auf jeweils einen hochdimensionalen Merkmalsvektor. Eine gängige Wahl für das automatisierte Topic Modeling war hier bisher die TF-IDF[19]-Repräsentation der Dokumente[20]. Bei dieser Form der Repräsentation werden die Worthäufigkeiten eines Dokuments, normiert bezüglich der Gesamthäufigkeit innerhalb der Sammlung, als die wesentlichen Merkmale betrachtet. Diese Merkmalsrepräsentation bringt eine Reihe von Nachteilen mit sich, so sind die resultierenden Merkmalsvektoren spärlich besetzt, die einzelnen Merkmale semantisch wenig aussagekräftig und insbesondere variable historische Schreibweisen sowie OCR-Fehler führen zu weiteren Problemen.

Nun betrachten wir eine Textsammlung, für die zuvor ein Entity-Linking berechnet wurde. Die beschreibenden Merkmale eines Dokuments, die für das Topic Modeling verwendet werden, entsprechen in diesem Fall der Summe der Verknüpfungswahrscheinlichkeiten pro Entität pro Dokument normiert bezüglich der Gesamtsumme dieser Wahrscheinlichkeiten über die gesamte Sammlung. Statt Wörtern werden also die Entitäten der Klassen Personen, Orte und Organisationen als die wesentlichen Eigenschaften, die den Inhalt eines Dokuments beschreiben, betrachtet. Diese Merkmale sind semantisch deutlich aussagekräftiger als Worthäufigkeiten und auch invarianter gegenüber unterschiedlichen Schreibweisen oder OCR-Fehlern. Dies lässt sich mit dem Einsatz von NER und EL begründen, mit deren Hilfe die sprachliche Variabilität zu einem gewissen Grad nivelliert wird. Gleichzeitig weisen die Merkmale aber auch Nachteile auf, z.B. werden Werke, die keine oder nur wenige eindeutig identifizierbare Entitäten erwähnen, unscharf abgebildet.

Topic Modeling ist ein aktives Forschungsfeld, in welchem eine breite Auswahl an algorithmischen Ansätzen für die automatisierte Bestimmung von Themengebieten existiert[21]. Im Rahmen einer Fallstudie wurde an der SBB die seit langem etablierte Latent-Dirichlet-Allocation-Methode (LDA)[22] genutzt, um die Werke der digitalisierten Sammlungen zu untersuchen. Dabei wurde die Implementation aus dem gensim[23] Python-Paket verwendet.

17 https://sonar.fh-potsdam.de/ [22.04.2024].

18 Vgl. Moretti, Franco: Conjectures on world literature, in: New Left Review 1 (2000). S. 54-68. Der Begriff des Distant Reading steht im Gegensatz zur herkömmlichen Einzeltextlektüre (Close Reading), welche in der Vergangenheit die vorherrschende geisteswissenschaftliche Arbeitsweise konstituierte.

19 Steht für: Term Frequency/Inverse Document Frequency, ein u.a. im Bereich des Information Retrieval weit verbreitetes Maß zur Relevanzbestimmung von Termen für einzelne Dokumente innerhalb eines Korpus.

20 Manning, Christopher / Schütze, Hinrich: Foundations of statistical natural language processing. MIT press, 1999.

21 Kherwa, Pooja / Bansal, Poonam: Topic modeling: a comprehensive review, in: EAI Endorsed transactions on scalable information systems 7, 24 (2019). https://eudl.eu/pdf/10.4108/eai.13-7-2018.159623.

22 Blei, David M. / Ng, Andrew Y. / Jordan, Michael I.: Latent Dirichlet Allocation, in: Journal of Machine Learning Research 3 (2003). S. 993-1022.

23 https://github.com/piskvorky/gensim [22.04.2024].

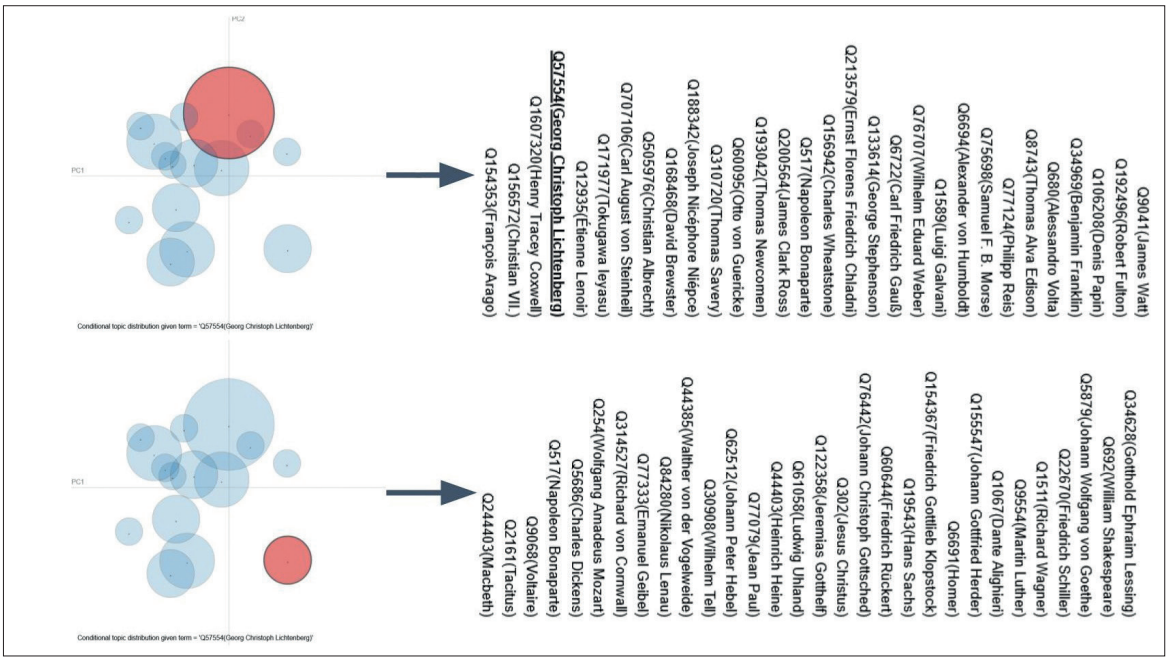

Abbildung 8: Topic Modeling der Digitalisierten Sammlungen auf Basis von Entity-Linking. Bedingte Themenverteilungen für die Entität „Georg Christoph Lichtenberg (Q57554)" ermöglichen die Identifikation verwandter Entitäten.

Für die Visualisierung wurde eine angepasste Version des ldavis[24] Javascript Paketes erstellt. Dafür wurden separate Themenmodelle für jeweils Personen, Orte und Organisationen getrennt und zusätzlich für Personen, Orte und Organisationen gemeinsam berechnet. Die Anzahl der Themenschwerpunkte, die zu bestimmen waren, wurde zwischen 70 und 280 Themengebieten variiert. Aus der Datenbasis der ca. 5 Millionen deutschsprachigen Seiten aus ca. 20.000 Werken wurden von den ca. 65 Millionen Verknüpfungen nur diejenigen mit einer Verknüpfungswahrscheinlichkeit größer als 0.5 für das Topic Modeling verwendet. Zudem wurden dabei nur Entitäten berücksichtigt, die in mindestens 0.5 Prozent der Werke Erwähnung finden.

Die angepasste ldavis-Benutzerschnittstelle[25] visualisiert die berechneten Themenverteilungen und erlaubt über die Selektion einzelner Themengebiete Einblicke in die Entitätenverteilung pro Gebiet, siehe auch Abbildung 8. Die Schnittstelle erlaubt das Umschalten zwischen Personen, Orten, Organisationen und einer Kombination davon sowie eine gestaffelte Auswahl der Anzahl der Themengebiete zwischen 70 und 280. Entitäten können ausgewählt oder gesucht werden, um resultierende bedingte Themenverteilungen anzuzeigen. Auch die jeweils zugehörigen Werke werden nach Relevanz gelistet angezeigt mit den entsprechenden Verweisen auf die semantisch angereicherten Textstellen aus den Digitalisierten Sammlungen. Ein Beispiel hierfür wird in Abbildung 9 gezeigt.

Abbildung 9: Links: Werke aus den Digitalisierten Sammlungen, die gemäß bedingter Themenverteilung für die Entität „Georg Christoph Lichtenberg (Q57554)" gruppiert wurden. Rechts: OCR-Textstelle aus dem „Regierungs- und Intelligenzblatt für das Herzogtum Gotha (1847)" mit Bezug zu Georg Christoph Lichtenberg.

Ressourcen

Wir streben an, sowohl relevante Datensätze als auch Quellcode von entwickelten Verfahren soweit rechtlich möglich online zur Verfügung zu stellen. Die Demonstratoren für die Entitätenerkennung und -verknüpfung[26] sowie eine Visualisierung des Topic Modeling[27] können online evaluiert werden. Diese sind explizit nicht als finale Produkte zu betrachten, sondern sollen als Machbarkeitsstudien verstanden werden, die Potentiale für die Weiterentwicklung der Digitalisierten Sammlungen aufzeigen. Durch Datenpublikationen sollen Forschungsgemein-

24 Sievert, Carson / Shirley, Kenneth: LDAvis: A method for visualizing and interpreting topics, in: Proceedings of the workshop on interactive language learning, visualization, and interfaces 2014. S. 63-70. https://aclanthology.org/W14-3110.pdf. S.a. https://github.com/cpsievert/LDAvis [22.04.2024].

25 Topic Modeling: https://ravius.sbb.berlin/sbb-tools/ldavis.html?&selected_map=Persons&n_topics=280 [22.4.2024].

26 NER+EL: https://ravius.sbb.berlin/sbb-tools/index.html?ppn=756689090 [22.4.2024].

27 Topic Modeling: https://ravius.sbb.berlin/sbb-tools/ldavis.html?&selected_map=Persons&n_topics=280 [22.4.2024].

schaften, beispielsweise der Bereiche Data Science, Digital Humanities oder Machine Learning, bei der Erschließung der Digitalisierten Sammlungen als Forschungsgegenstand unterstützt werden. Insbesondere die Wahl von Dateiformaten, die durch die Forschenden unkompliziert und direkt genutzt werden können, steht dabei im Vordergrund. Alle extrahierten Volltexte der Sammlung wurden in Form von SQLite-Datenbankdateien publiziert[28]. Auch die vorberechnete Wissensbasis aus Wikidata und Wikipedia sowie die resultierende semantische Anreicherung durch NER und EL der Digitalisierten Sammlungen ist als SQLite-Datenbank online abrufbar[29]. Der Quellcode für Entitätenerkennung[30] und -verknüpfung[31], Extraktion der Wissensbasis aus Wikidata und Wikipedia[32] sowie Berechnung und Visualisierung des Topic Modeling[33] steht online zur Nachnutzung bereit.

Ausblick

Die hier vorgestellten, an der Staatsbibliothek zu Berlin entwickelten Systeme für Named Entity Recognition und Linking wurden in entsprechenden Benchmarks insbesondere im Hinblick auf die Erkennung und Verlinkung von Entitäten in historischen und mehrsprachigen Dokumenten evaluiert[34], die Ergebnisse deuten auf eine robuste Performance der Systeme sowie auf eine vielseitige Einsetzbarkeit für heterogene Materialien hin. Vor dem Hintergrund der großen Varianz der Bestände in den Digitalisierten Sammlungen und den noch vorhandenen Herausforderungen durch historische Sprachvarianten und OCR-Fehler zeigt NER/EL damit eine hinreichende Reife bei gleichzeitig vielfältigen Potentialen, so dass die Anreicherung von Dokumenten mit NER/EL zusätzlich zur OCR-Verarbeitung als Standard für alle Materialien in die Digitalisierungsprozesse integriert werden sollte.

Gleichzeitig eröffnen rasante Entwicklungen auf dem Feld des Machine- und Deep Learning umfangreiches Weiterentwicklungs- und Optimierungspotenzial. Ein guter Startpunkt sind außerdem unsere Daten: während die stetige Ausdehnung sowie Korrektur der Ground Truth Datenbasis auf lange Sicht maßgeblich zu besseren Ergebnissen beiträgt, sehen wir weitere Möglichkeiten in der Schaffung von Interoperabilität zwischen unseren und anderen auf historisches NER/EL spezialisierten Datensätzen.

Gleichermaßen haben wir aufgezeigt, von welchen Vorteilen die Nutzenden bei der Suche, Exploration und Untersuchung von Beständen sowie den darin genannten Entitäten profitieren können, wenn eine Einbindung der NER-/EL-Ergebnisse z.B. in die Digitalisierten Sammlungen der SBB gegeben ist. Auch eine Integration mit weiteren Diensten wie beispielsweise der Bildähnlichkeitssuche, die aktuell im Mensch.Maschine.Kultur Projekt entwickelt wird, ist denkbar. So könnte beispielsweise der Aufruf einer in einem Text der Digitalisierten Sammlungen enthaltenen und mittels NER identifizierten Entität eine (textuelle) Suche innerhalb der Bildsuche starten. Umgekehrt können in der Bildähnlichkeitssuche auch zusätzliche Suchfacetten implementiert werden, die eine Eingrenzung vorhandener Bilder nach damit verwandten Entitäten über ihre Relation zu Text(-seiten) der Digitalisierten Sammlungen erlauben[35]. NER in Kombination mit anderen im Projekt entwickelten Technologien wie der Layout- oder Texterkennung kann zudem dazu genutzt werden, weitere strukturierte Informationen aus Bildern zu extrahieren, die wiederum als Trainingsdaten bzw. zur Erweiterung bestehender Trainingsdaten für die Bildähnlichkeitssuche verwendet werden können[36]. Nicht zuletzt können die mittels Named Entity Recognition und Linking extrahierten inhaltlichen Metadaten auch lange etablierte Dienstleistungen der Bibliothek wie die Erschließung[37] oder Entwicklung von Suchinstrumenten[38] weiter bereichern. Neben den Entitäten selbst lassen sich über das Linking zu einer Wissensbasis wie Wikidata auch ganz neuartige Informationen abrufen und anzeigen, beispielsweise biografische Daten zu Personen, sowie semantische Relationen zwischen Entitäten identifizieren[39].

28 Volltexte der Digitalisierten Sammlungen: https://zenodo.org/records/7716098 [22.4.2022].

29 NER+EL Daten: https://zenodo.org/records/7767404 [deutsch]. NER+EL Daten: https://zenodo.org/records/7773746 [französisch]. NER+EL Daten: https://zenodo.org/records/7773987 [englisch].

30 Quellcode NER: https://github.com/qurator-spk/sbb_ner [22.4.2024]

31 Quellcode EL: https://github.com/qurator-spk/sbb_ned

32 Quellcode Extraktion Wikidata/Wikipedia: https://github.com/qurator-spk/sbb_knowledge-base

33 Quellcode Topic Modeling: https://github.com/qurator-spk/sbb_topic-modelling

34 https://impresso.github.io/CLEF-HIPE-2020/ [22.04.2024] und https://hipe-eval.github.io/HIPE-2022/ [22.04.2024].

35 Vgl. z.B. Lin, Yiling / Ahn, Jae-Wook / Brusilovsky, Peter / He, Daqing / Real, William: Imagesieve: Exploratory search of museum archives with named entity-based faceted browsing. Proceedings of the American Society for Information Science and Technology 47, 1 (2010). S. 1-10. https://doi.org/10.1002/meet.14504701217.

36 Takano, Atsuko / Cole, Theodor C. H. / Hajime Konagai: A novel automated label data extraction and data base generation system from herbarium specimen images using OCR and NER, in: Sci Rep 14, 112 (2024). https://doi.org/10.1038/s41598-023-50179-0.

37 Vgl. z.B. Goh, Rachael: Using named entity recognition for automatic indexing, in: IFLA Proceedings 2017. https://library.ifla.org/id/eprint/2214/1/115-goh-en.pdf.

38 Lizarralde, Ignacio / Mateos, Cristian / Rodriguez, Juan Manuel / Zunino, Alejandro: Exploiting named entity recognition for improving syntactic-based web service discovery, in: Journal of Information Science 45, 3 2019. S. 398-415. https://doi.org/10.1177/0165551518793321.

39 Düring, Marten / Bunout, Estelle / Guido, Daniele: Transparent Generosity: Introducing the impresso interface for the exploration of semantically enriched historical newspapers, 2022. https://hal.science/hal-04154431/.

Sophie Schneider
ist seit Dezember 2022 Wissenschaftliche Mitarbeiterin im Forschungsprojekt „Mensch.Maschine.Kultur“ an der SBB und zuständig für den Teilbereich zur KI-unterstützten Inhaltsanalyse und Sacherschließung. Sie studierte Bibliotheks- und Informationswissenschaft in Potsdam und Berlin und interessiert sich für die Themen Digital Humanities, Natural Language Processing und Maschinelles Lernen.
ORCID: 0000-0002-8303-1798
bibwiss.github.io
sophie.schneider@sbb.spk-berlin.de

Clemens Neudecker
Studium der Philosophie, Informatik und Politischen Wissenschaften, arbeitet als Referatsleiter Data Science in der Abteilung Informations- und Datenmanagement der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz. Seit beinahe 20 Jahren richtet sich sein Forschungsinteresse vor allem auf die Bereiche Computer Vision und Natural Language Processing sowie deren Anwendung im Kontext von Digitalisierung und den Digital Humanities.
ORCID: 0000-0001-5293-8322
https://cneud.net

Kai Labusch
arbeitet als wissenschaftlicher Mitarbeiter im Projekt „Mensch.Maschine.Kultur“ der Staatsbibliothek zu Berlin. Er studierte Informatik an der Universität zu Lübeck mit den Nebenfächern Neuroinformatik und Biomathematik und promovierte zum Thema „Soft-competitive learning of sparse data models“. 2011 wechselte er in die Softwarebranche, wo er Methoden des maschinellen Lernens auf NLP-Probleme, Prozessoptimierung in der Stahlindustrie und Visualisierung anwandte. 2019 kam er als Teil des Qurator-Teams zur Staatsbibliothek zu Berlin, wo er sich auf Entitätserkennung und -verknüpfung sowie Bildsuchanwendungen konzentrierte.
ORCID: 0000-0002-7275-5483
Kai.Labusch@sbb.spk-berlin.de