Wenn Algorithmen Zeitschriften lesen. Vom Mehrwert automatisierter Textanreicherung

Michael Gasser, ETH Zürich, ETH-Bibliothek, Archive
Regina Wanger, ETH Zürich, ETH-Bibliothek, DigiCenter
Ismail Prada, Universität Zürich, Institut für Computerlinguistik

Zusammenfassung:
In Zusammenarbeit mit dem Institut für Computerlinguistik der Universität Zürich (ICL UZH) lancierte die ETH-Bibliothek Zürich ein Pilotprojekt im Bereich automatisierter Textanreicherung. Grundlage für den Piloten bildeten Volltextdateien der Schweizer Zeitschriftenplattform E-Periodica. Anhand eines ausgewählten Korpus dieser OCR-Daten wurden mit automatisierten Verfahren Tests in den Bereichen OCR-Korrektur, Erkennung von Personen-, Orts- und Ländernamen sowie Verlinkung identifizierter Personen mit der Gemeinsamen Normdatei GND durchgeführt. Insgesamt wurden sehr positive Resultate erzielt. Das verwendete System dient nun als Grundlage für den weiteren Kompetenzausbau der ETH-Bibliothek auf diesem Gebiet. Das gesamte bestehende Angebot der Plattform E-Periodica soll automatisiert angereichert und um neue Funktionalitäten erweitert werden. Dies mit dem Ziel, Forschenden einen Mehrwert bei der Informationsbeschaffung zu bieten. Im vorliegenden Beitrag werden Projektinhalt, Methodik und Resultate erläutert sowie das weitere Vorgehen skizziert.

Summary:
In cooperation with the Institute of Computational Linguistics at the University of Zurich (ICL UZH), the ETH Library Zurich carried out a pilot project in the field of automated text enrichment. The basis for the pilot were full text files from E-Periodica, the online platform for digitised Swiss journals. Based on a selected corpus of this OCR data and using automated procedures, tests were performed in the areas of OCR correction, recognition of person, place and country names as well as linking identified persons to the German common authority file for libraries (GND). Overall, very positive results were achieved. The system used now serves as a basis for the further expansion of the ETH Library's competence in this field. The entire content of the E-Periodica platform is to be automatically enhanced and extended with new functionalities. The aim is to offer researchers added value in information gathering. In this article, project content, methodology and results are presented and the next steps are outlined.

Zitierfähiger Link (DOI); https://doi.org/10.5282/o-bib/2018H4S181-192
Autorenidentifikation: Gasser, Michael: GND 1120320976,
ORCID: https://orcid.org/0000-0003-0390-1448;
Prada, Ismail: ORCID: https://orcid.org/0000-0003-4229-8688;
Wanger, Regina: GND 1172079137
Schlagwörter: Bibliothekswesen; Computerlinguistik; automatisierte Textanreicherung; Named Entity Recognition (NER); Named Entity Linking (NEL); OCR-Optimierung

1. E-Periodica – Millionen von Zeitschriftenseiten im Volltext

Seit 2016 betreibt die ETH-Bibliothek Zürich E-Periodica, die Plattform für digitalisierte Schweizer Zeitschriften.[1] E-Periodica bietet freien Zugang zu den Volltexten von über 300 Zeitschriften. Thematisch decken die angebotenen Titel von Natur- und Ingenieurwissenschaften über Architektur und Kunst bis zu Geschichte, Geographie und Religion ein breites Spektrum ab. Zeitlich liegt der Schwerpunkt auf aktuellen Zeitschriften und Titeln des 19. und 20. Jahrhunderts. Die ältesten Inhalte reichen aber bis ins 18. Jahrhundert zurück. Das Angebot der Plattform wird laufend erweitert. Dies sowohl mit der Aufschaltung aktueller Nummern als auch mit der Integration zusätzlicher Zeitschriften, die i. d. R. im hauseigenen DigiCenter der ETH-Bibliothek retrodigitalisiert werden.

Die OCR-Erkennung ist Grundlage für die Volltextsuche, die in den insgesamt über 7 Millionen auf E-Periodica aufgeschalteten Seiten möglich ist. Diese OCR-Erkennung wird mit der Software ABBYY FineReader durchgeführt. Die Volltexte werden zudem in Form von PDFs zur Verfügung gestellt, die Benutzende von E-Periodica beispielsweise von einzelnen Artikeln herunterladen können. Die Volltextdateien bilden zusammen mit den TIFF-Files und den in Form von XML-Dateien vorliegenden Metadaten den Datenbestand des Online-Angebots, das 2016 über 7 Millionen Zugriffe verzeichnete.[2]

2. Pilotprojekt zur Eigennamenerkennung mit vierfachem Mehrwert

Es erstaunt nicht, dass dieses beachtliche Textkorpus von E-Periodica eine wichtige Rolle spielte, als im Herbst 2016 die in der ETH-Bibliothek neu gegründete Arbeitsgruppe Text- und Datamining daran ging, mögliche Handlungsfelder für ein Pilotprojekt zu definieren. Unter den vorgeschlagenen und diskutierten Themenvorschlägen kristallisierte sich als priorisiertes Feld rasch ein Projekt auf dem Gebiet der automatisierten Textanreicherung heraus. Die grosse Menge an bereits vorhandenen OCR-Dateien in E-Periodica stellte für diesen Zweck eine geradezu ideale Ausgangsbasis dar.

Das Pilotprojekt, das schwerpunktmässig in der zweiten Jahreshälfte 2017 auf der Datenbasis eines Samples zweier Architekturzeitschriften aus E-Periodica umgesetzt wurde, hatte zum Ziel, Mehrwert in vier Bereichen zu schaffen:

- Anreicherung der vorhandenen Volltextdateien durch automatisierte Eigennamenerkennung (Personen, Orte, Länder) in Kooperation mit einem interessierten Forschungspartner
- Kompetenzgewinn und Ausbau des vorhandenen Fachwissens innerhalb der ETH-Bibliothek
- Die Entwicklung neuer und kundenorientierter Zusatzfunktionen in E-Periodica auf Basis der angereicherten Daten
- Die Definition von Fragen für weiterführende Untersuchungen sowohl in der Forschung als auch in der praktischen Anwendung

1 E-Periodica (<www.e-periodica.ch>) löste 2016 die seit 2006 bestehende Vorgängerplattform retro.seals.ch ab.
2 ETH-Bibliothek Zürich (Hg.): ETH-Bibliothek Jahresbericht 2016, Zürich 2017, S. 66. Online: <https://doi.org/10.3929/ethz-a-004157606>.

Aufgrund der inhaltlichen Breite des Vorhabens wurden innerhalb der Bibliothek nicht nur der Product Owner bzw. der IT-Verantwortliche der Plattform E-Periodica in das Projektteam eingebunden. Hinzu kamen auch interne Fachleute aus den Bibliotheks-IT-Services, der Innovation und den Informationswissenschaften. Von Anfang an stand jedoch fest, dass das Pilotprojekt nur in engem Austausch mit Benutzenden und externen Partnern erfolgreich sein konnte. Im Vordergrund stand dabei primär die grundlegende Kooperation mit einem interessierten Forschungspartner aus dem Bereich der Computerlinguistik.

3. Automatisierte Eigennamenerkennung

3.1. Forschungspartner und Ziele

Rasch konnte für das Pilotprojekt das Institut für Computerlinguistik der Universität Zürich (ICL UZH) als kompetenten Kooperationspartner gewonnen werden. Einer der Forschungsschwerpunkte des Instituts unter der Leitung von Prof. Martin Volk liegt im Bereich der Digital Humanities. Entsprechend verfügt es über breite Erfahrung im Umgang mit Korpora retrodigitalisierter Zeitschriften und Zeitungen.

Die Tätigkeit des Instituts für Computerlinguistik konzentrierte sich im Rahmen des fünfmonatigen Pilotprojektes (Juli bis November 2017) auf folgende Ziele:

- Experimente und Evaluation zur muster-basierten Korrektur von OCR-Fehlern
- Anpassung und Evaluation des institutseigenen, regelbasierten Systems zur Erkennung von Personennamen in deutschsprachigen Texten, das ursprünglich für die Namenserkennung in alpinistischen Texten entwickelt wurde[3]
- Experimente zur Verlinkung der erkannten Personennamen mit der Gemeinsamen Normdatei (GND)
- Anpassung und Evaluation des institutseigenen Systems zur Erkennung von Ländernamen und Schweizer Ortsnamen in deutschsprachigen Texten
- Vergleich des institutseigenen Systems mit einem anderen statistischen System

Hand in Hand mit der kurzen Laufzeit des Pilotprojektes ging der bewusste Einsatz eines bewährten, regelbasierten Verfahrens. Im Vordergrund der Kooperation mit dem ICL UZH stand die Übertragung dessen vorhandenen Systems auf den E-Periodica-Korpus und v. a. die Ergänzung um die bisher nicht vorhandene automatisierte GND-Verlinkung. Das Vorgehen und die erzielten Ergebnisse wurden seitens des ICL UZH von Prof. Martin Volk und seinem Projektmitarbeiter Ismail Prada für die ETH-Bibliothek in einem internen Projektbericht festgehalten.[4] Auf diesem Dokument basieren die hier publizierten Ausführungen.

3 Ebling, S; Sennrich, R; Klaper, D; Volk, Martin: Digging for names in the mountains: Combined person name recognition and reference resolution for German alpine texts, in: 5th Language & Technology Conference, Poznan, Poland, 25 November 2011 - 27 November 2011. Online: <https://doi.org/10.5167/uzh-50451>. Siehe auch Universität Zürich: Text+Berg digital, <http://textberg.ch>, Stand 24.09.2018.

4 Volk, Martin; Prada, Ismail: Bericht zum Eigennamenerkennung-Pilotprojekt, Zürich Dezember 2017 (unveröffentlichter Projektbericht).

3.2. Datenbasis und Goldstandard

Aus dem ganzen E-Periodica-Angebot wurden für das Pilotprojekt zwei umfangreiche Schweizer Architekturzeitschriften als Datenbasis ausgewählt:

- *Tec21* (bis 1978 *Schweizerische Bauzeitung*), 142 Jahrgänge (1874–2016), ca. 270‘000 Seiten
- *Werk, Bauen und Wohnen*, 102 Jahrgänge (1914–2016), ca. 112‘000 Seiten

In Absprache mit den jeweiligen Verlagen wurden dem ICL UZH im Rahmen des Pilotprojekts über 380‘000 OCR-Textdateien als Korpus zur Verfügung gestellt.

Als Referenz für die qualitative Bewertung der Ergebnisse der automatischen OCR-Verbesserung und Eigennamenerkennung durch das verwendete System wurde anhand dreier *Tec21*-Jahrgänge ein sogenannter Goldstandard, ein manuell annotiertes und korrigiertes Subset des Korpus, erstellt. Dabei wurden für die Jahrgänge 1895, 1940 und 1990 der Zeitschrift *Tec21* sämtliche relevanten Zielinformationen erfasst (s. Tabelle 1).

Tabelle 1: Werte des manuell annotierten Goldstandards im Überblick

	1895	1940	1990
Wörter OCR-Text	6971	14403	7911
Seiten OCR-Text	7	12	9
Wörter mit OCR-Fehlern	164	119	15
Wörter in Personennamen	153	269	200
Schweizer Ortsnamen	16	55	13
Ländernamen	20	12	10

3.3. Automatische OCR-Korrektur

Generell wurde die Qualität der OCR in den gelieferten Texten als sehr gut beurteilt. Umso mehr wurde im Pilotprojekt grosser Wert darauf gelegt, automatische OCR-Korrekturen nur mit hoher Präzision vorzunehmen. Es sollten also nur Wörter automatisch ersetzt werden, die mit hoher Sicherheit OCR-Fehler enthielten und für die mit ebenfalls hoher Gewissheit die korrekte Entsprechung ermittelt werden konnte. Die Zahl der Falschkorrekturen (*false positives*) sollte möglichst tief gehalten werden.

Um dieses Ziel im gegebenen Zeitrahmen zu erreichen, wurde mit verschiedenen Beschränkungen gearbeitet. So wurde etwa für jeden Satz festgestellt, in welcher Sprache er geschrieben ist. Nur Wörter in deutschen Sätzen wurden überhaupt in die Korrektur miteinbezogen. Bei der Identifikation

der einzelnen Wörter mithilfe des TreeTaggers[5] wurden ausschliesslich Wörter mit einer Zeichenlänge von mindestens sechs Zeichen berücksichtigt. Bei einem nicht identifizierten Lemma wurde in dessen Umfeld – drei Seiten vorher bis drei Seiten nachher – mittels Ähnlichkeitsmessung nach dem korrekten bekannten Begriff gesucht, um dann die Ersetzung vorzunehmen. Auf diese Weise konnte etwa die fehlerhafte OCR-Erkennung „Haiienstadion" erfolgreich in „Hallenstadion" korrigiert werden.

Der Abgleich mit dem Goldstandard zeigt, dass durch diese Verfahren die Präzision, d. h. die korrekt korrigierten OCR-Fehler, tatsächlich sehr hoch gehalten werden konnte. Für die Jahrgänge 1895 und 1990 konnte sogar eine Präzision von 100 % erreicht werden. Im Jahrgang 1940 wurde nur eine einzige Falschkorrektur vorgenommen. Allerdings ging diese hohe Präzision sehr deutlich zulasten einer hohen Ausbeute (s. Abb. 1). Die grosse Mehrheit der vorhandenen OCR-Fehler wurde nicht als solche erkannt bzw. nicht korrigiert. Indem dies jedoch eine Folge des zeitlich eng limitierten Pilotprojektes mit Experimentcharakter ist, kann davon ausgegangen werden, dass im Bereich der OCR-Nachbearbeitung die Optimierungsmöglichkeiten bei weitem noch nicht ausgeschöpft sind.

3.4. Orts- und Ländernamenerkennung

Für das Pilotprojekt wurde die Erkennung von Ortsnamen auf die Schweiz eingeschränkt. Das gewählte Verfahren beruhte dabei – wie auch bei der Erkennung von Ländernamen – auf einem Listenabgleich. Lediglich einige mehrdeutige Bezeichnungen wurden aus der Liste ausgeschlossen. Das führte dazu, dass z. B. die Städte „Baden" oder „Zug" nicht gefunden werden konnten, weil sich diese auf „Baden" im Sinne von Schwimmen bzw. „Zug" im Sinne von Eisenbahn beziehen könnten. Bei den Ländern wurde eine Liste verwendet, die auch Namen inzwischen nicht mehr existierender Staaten des 19. und 20. Jahrhunderts enthält. Einzelne Schreibvarianten von Abkürzungen (z. B. „USA", „U.S.A", „U. S. A.") wurden in der Liste nachgetragen. Um das System künftig weiter zu verbessern und weitere Varianten von Ort- und Ländernamen erkennen zu können, liegt es nahe, die Liste mit Namensvarianten aus zusätzlichen Datenquellen (etwa der GND) zu ergänzen.

Aber auch so wurden mit dem angewandten Verfahren sowohl bei der Länder- als auch bei der Ortsnamenerkennung sehr gute Resultate erzielt. Gemessen am Goldstandard wurde bei hoher Präzision eine hohe Ausbeute erzielt (s. Abb. 1). Generell gilt dabei: Je höher die Präzision desto tiefer die Rate von Falschkorrekturen und je höher die Ausbeute desto tiefer die Zahl fälschlicherweise übersehener OCR-Fehler bzw. Namen. Bei den Ländern ist die hohe Präzision und Ausbeute nicht zuletzt darauf zurückzuführen, dass es kaum Ländernamen gibt, die daneben auch andere Bedeutungen haben. Bei den Ortsnamen hingegen ist Ambiguität eine grössere Fehlerquelle. So ist etwa nicht eindeutig, ob mit „Beznau" die Ortschaft oder das gleichnamige Kernkraftwerk gemeint ist.

5 Schmid, Helmut: TreeTagger – a part-of-speech tagger for many languages, <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>, Stand: 24.09.2018. Der TreeTagger ist ein System, welches zu jedem Wort Lemma und Wortart bestimmt. Wenn er für ein Wort kein Lemma bestimmen kann, kann davon ausgegangen werden, dass es sich dabei nicht um ein gültiges Wort der deutschen Sprache handelt – oder eben um einen Eigennamen.

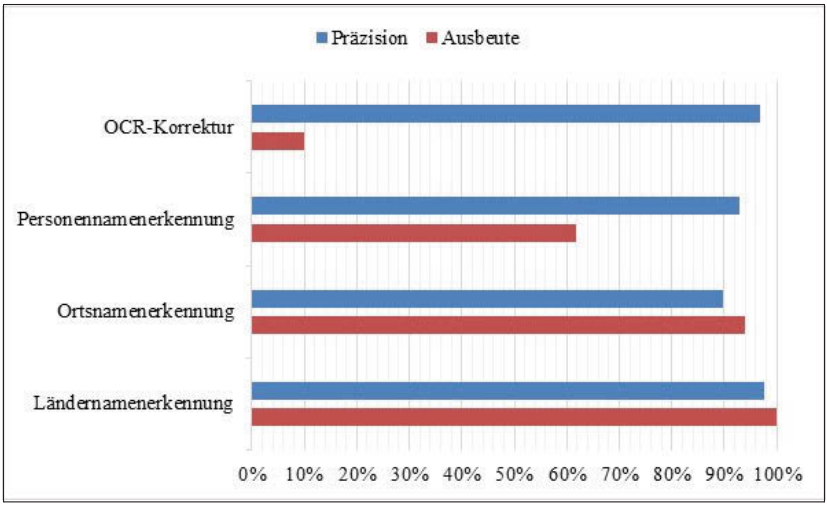

Abb.1: Resultate der Präzision bzw. Ausbeute der OCR-Korrektur sowie der Personen-, Länder- und Ortsnamenerkennung gemessen am Goldstandard

Die identifizierten Orts- bzw. Ländernamen wurden pro Zeitschrift und Jahr in separaten XML-Ausgabedateien festgehalten. Diese Dateien enthalten nicht nur die Namen der jeweiligen Entitäten mit den genauen Positionsangaben, an denen sie gefunden wurden. Ausgegeben wurden ebenfalls die entsprechenden Postleitzahlen der Schweizer Ortschaften bzw. der numerische ISO-3166-Code der identifizierten Länder. Auf dieser Basis können Verknüpfungen sowohl innerhalb des Korpus (z. B. zwischen verschiedenen Schreibvarianten eines Landes) als auch zu externen Datenquellen vorgenommen werden.

3.5. Personennamenerkennung

Vergleichbar mit der Erkennung von Orts- und Personennamen setzte das ICL UZH im Rahmen dieses Pilotprojektes auch bei der Personenerkennung ein listen- und regelbasiertes Verfahren ein. In diesem Fall war der Ausgangspunkt eine umfangreiche Liste von Vornamen. Die Erkennung und das Lernen von Nachnamen beruhten dann auf allgemeinen Regeln. Dazu gehört z. B., dass nur Worte, die mit einem Grossbuchstaben beginnen, überhaupt als Nachnamen infrage kommen. Andere Regeln basieren auf Treffern in der Vornamenliste: Ein Wort nach einem Vornamen ist ein Nachname („Robert Maillart"). Ein dritter Teil des Regelsets bezog den Kontext des jeweiligen Zeitschriftenjahrgangs mit ein: Ein Wort, das im selben Jahrgang an anderer Stelle als Nachname identifiziert wurde, ist ein Nachname (Identifikation von „Robert Maillart" ermöglicht im selben Jahrgang die Personennamenerkennung in „die Brücke von Maillart"). Schliesslich wurden auch erkannte Titel und Berufsbezeichnungen als Marker genutzt: Ein Begriff nach einer Berufsbezeichnung ist dann ein Nachname, wenn es kein reguläres Wort ist, das mithilfe des TreeTaggers bestimmt werden konnte („Bauingenieur Maillart"). Für die Ausgabedatei wurden pro Zeitschrift und Jahrgang die gefundenen Merkmale identifizierter Personen aggregiert. Im besten Fall konnten für eine Person folgende Informationen aus dem Korpus ausgelesen werden: Name, Vorname (inkl. Abkürzungen), Geburtsjahr, Geschlecht und Beruf.

Mit Blick auf eine künftige Systemoptimierung bietet es sich gerade bei der verwendeten Liste von Berufen an, diese durch die in der GND verwendeten Bezeichnungen zu ergänzen.

Auch bei der automatisierten Personennamenerkennung waren eine möglichst hohe Präzision und damit die Vermeidung von *false positives* das Ziel. Gemessen am Goldstandard wurde mit dem verwendeten System tatsächlich eine sehr zuverlässige Erkennung mit einer Präzision von über 90 % erreicht. Nur eine geringe Zahl von Wörtern wurde fälschlicherweise als Namen identifiziert. Probleme verursachten in erster Linie Konstruktionen wie „Charles Aussage" oder „T. H. Stuttgart", eine Abkürzung für „Technische Hochschule Stuttgart". Die hohe Präzision wirkte sich allerdings auch hier deutlich auf die Ausbeute aus (s. Abb. 1). So wurde im Jahrgang 1990 rund ein Viertel, im Jahrgang 1940 fast die Hälfte aller Namen nicht erkannt.

In ganz kleinem Rahmen wurden im Pilotprojekt die Resultate des regelbasierten Systems des ICL UZH mit dem statistischen Verfahren der Zürcher Hochschule für Angewandte Wissenschaft (ZHAW), School of Engineering, verglichen.[6] Der Vergleich beschränkte sich lediglich auf die Erkennung von Personennamen in den drei Zeitschriftenjahrgängen des Goldstandards. Die Ergebnisse lagen im Bereich des Erwarteten: Das statistische System liefert eine höhere Ausbeute an Personennamen, wies aber gleichzeitig eine niedrigere Präzision aus. Ein naheliegender Schluss ist, die Ergebnisse des regelbasierten Systems für das Training des statistischen Verfahrens zu verwenden, sofern die Ergebnisse für eine einzelne Zeitschrift oder ein klar definiertes Set an ausgewählten Zeitschriften optimiert werden sollen.

3.6. Verlinkung mit der GND

Als Versuch einer zusätzlichen Anreicherung des Textkorpus verlinkte das ICL UZH die identifizierten Personen automatisiert mit der Gemeinsamen Normdatei (GND).

Das System nutzte dafür die aggregierten Informationen aus der Personennamenerkennung zum Abgleich mit der GND, die im MARC21-XML-Format zur Verfügung stand. Stimmten sämtliche fünf Merkmale (Name, Vorname, Geburtsjahr, Geschlecht und Beruf) mit der GND überein, wurde der ausgelesenen GND-ID im zusätzlichen Attribut <gnd_certainty> der höchstmögliche Wert „5" zugewiesen. Erfolgte die Zuweisung aufgrund von vier Merkmalen oder weniger, sank dieser Wert entsprechend auf „4", „3", „2" oder gar „1". Damit ein Eintrag in der GND als Kandidat für eine Verlinkung für einen erkannten Namen gilt, müssen der Nachname, wenigstens der abgekürzte Vorname sowie das Geschlecht übereinstimmen. Falls kein Geschlecht in der GND genannt wird, wird angenommen, dass das Geschlecht stimmt. Aus Plausibilitätsgründen muss die Zeitschrift mindestens 20 Jahre nach dem Geburtsdatum erschienen sein. Für die volle Punktzahl müssen ausserdem wenigstens ein ausgeschriebener Vorname sowie ein Beruf übereinstimmen. Einen Abzug gibt es, wenn die Berufe im

6 Statistische – oder auch maschinelle – Verfahren verwenden keine von Menschen geschriebenen Regeln zum Auffinden der Namen, sondern „Lernen" anhand grosser Mengen von getaggten Texten, wie Namen am besten zu identifizieren sind. Zum Verfahren der ZHAW siehe: von Däniken, Pius; Cieliebak, Mark: Transfer Learning and Sentence Level Features for Named Entity Recognition on Tweets, in: The Association for Computational Linguistics (Hg.), Proceedings of the 3rd Workshop on Noisy User-generated Text, Copenhagen, Denmark, September 7, 2017, S. 166-171. Online: <http://www.aclweb.org/anthology/W17-4422>, Stand: 24.09.2018.

GND-Eintrag zwar keine genaue Übereinstimmung zum erkannten Namen aufweisen, aber es sich dabei um wenigstens einen Beruf handelt, der typisch für Bauzeitschriften ist.[7] Beispiel: Ein GND-Eintrag, bei welchem Nachname, abgekürzter Vorname und Geschlecht übereinstimmen, und als Beruf „Ingenieur" eingetragen ist, dieser Beruf aber nicht im Text gefunden werden konnte, erhält nach diesem System eine Punktzahl von 3.

Ein Blick auf die Resultate für den Jahrgang 1940 der *Schweizerischen Bauzeitung* aus dem Goldstandard zeigt, dass lediglich einer relativ geringen Anzahl Personen mit höchster Zuverlässigkeit die entsprechende GND-ID zugewiesen werden konnte (s. Tabelle 2).

Tabelle 2: Korrekte Verlinkungen nach Sicherheit im Jahrgang 1940 der Schweizerischen Bauzeitung

GND Certainty	Korrekte Verlinkung	Total Verlinkungen
Stufe 5	92.3 %	12
Stufe 4	95.0 %	94
Stufe 3	60.0 %	123
Stufe 2	60.0 %	207
Stufe 1	65.0 %	251

Der starke Abfall korrekter Verlinkungen von Stufe 4 auf 3 ist ein deutlicher Hinweis darauf, dass lediglich die beiden höchsten Stufen für eine Verwertung der automatischen GND-Verlinkung infrage kommen. Auf diesen Stufen war auch das Problem mehrerer übereinstimmender GND-Einträge auf wenige Fälle beschränkt. Es erstaunt jedoch nicht, dass im Umfeld der verwendeten Schweizer Architekturzeitschriften in erster Linie prominentere Namen erfolgreich, eindeutig und mit hohem Sicherheitswert mit der GND verknüpft werden konnten. So wurde im Jahrgang 1940 z. B. der finnische Architekt Alvar Aalto und der Schweizer Landschaftsarchitekt Gustav Ammann mit dem höchsten Sicherheitswert verlinkt. Unter den Verlinkungen mit der zweithöchsten Stufe finden sich u. a. der Architekt und ETH-Professor Karl Moser oder der Bauingenieur, Brückenbauer und Unternehmer Robert Maillart.

Innerhalb der kurzen Laufzeit des Pilotprojekts konnten mögliche Optimierungen der GND-Verlinkung nicht mehr vorgenommen werden. Verbesserungsbedarf wurde z. B. beim Abgleich von Berufen erkannt. So sollten etwa „Ingenieur" und „Bauingenieur" als übereinstimmend angesehen werden.

4. Kompetenzgewinn der ETH-Bibliothek

Detaillierte Systemoptimierungen in den verschiedenen Bereichen der Eigennamenerkennung waren nicht das primäre Ziel des Pilotprojektes. Die erzielten Ergebnisse des ICL UZH zeigten jedoch klar, dass im E-Periodica-Textkorpus schon allein aufgrund regelbasierter Verfahren erfolgreich eine

7 Z. B. Architekt, Ingenieur, Mathematiker, u. ä. Die Liste solcher Berufe wurde jeweils anhand der gefundenen Berufe innerhalb des Jahrgangs generiert.

Eigennamenerkennung und Textanreicherung vorgenommen werden konnte. Umso wichtiger war es, im nächsten Schritt das im Pilotprojekt verwendete System an die ETH-Bibliothek zu transferieren. Der entsprechende Know-how-Transfer an die Bibliothek war Teil der mit dem Forschungsinstitut getroffenen Vereinbarung.

Dass sich unter den Mitarbeitenden der Bibliotheks-IT-Services bereits Fachleute mit computerlinguistischer Ausbildung befinden, erleichterte die Installation und Inbetriebnahme der gelieferten Skripte und Python-Programme erheblich. Seit Frühjahr 2018 werden auf dieser Basis v. a. in zwei Bereichen weitere Tests durchgeführt. Zum einen wird an der Verbesserung der OCR-Nachbearbeitung gearbeitet. Hier gilt es zu ermitteln, mit welchen Anpassungen – und natürlich ohne markante Einbusse bei der Präzision – die geringe Ausbeute erhöht werden kann, die im Pilotprojekt erzielt worden war. Wichtig ist dabei, möglichst nur generische Anpassungen vorzunehmen. Schliesslich soll die OCR-Verbesserung ohne den Aufwand spezifischer, u. U. von Zeitschrift zu Zeitschrift wechselnder Einstellungen für das ganze Spektrum an Zeitschriften in E-Periodica genutzt werden können.

Auf dieses Ziel ist auch die zweite Reihe von Tests ausgerichtet. Anhand von Testläufen mit zusätzlichen Zeitschriftentiteln wird das Verhältnis zwischen Präzision und Ausbeute bei der Eigennamenerkennung überprüft. Zunächst stehen dabei weitere Schweizer Architekturzeitschriften im Vordergrund. Anschliessend geht es um die Frage der Übertragbarkeit der Resultate auf Zeitschriften aus anderen Fachgebieten.

Das durchgeführte Pilotprojekt brachte und bringt der ETH-Bibliothek aber nicht nur einen Kompetenzgewinn im technisch-computerlinguistischen Bereich. Wichtig sind auch die Erfahrungen, die im Bereich der Rechteabklärungen gemacht wurden. Die vor dem Pilotprojekt mit Verlagen getroffene Vereinbarung zur freien Online-Publikation von Zeitschriften auf E-Periodica beschränkte sich allein auf die menschliche Nutzung. Computergestützte Analysen und Bearbeitungen der Volltexte waren darin nicht abgedeckt. Auf entsprechende Anfrage durch die ETH-Bibliothek hin gaben aber die Verlage der beiden im Pilotprojekt verwendeten Architekturzeitschriften, espazium Verlag und Verlag Werk AG, sehr schnell und bereitwillig ihre notwendige Zustimmung. Neu abgeschlossene Verträge enthalten nun eine diesbezüglich ergänzte Klausel.

5. Zusatzfunktionen für die User von E-Periodica

Von Beginn an war das Pilotprojekt auch darauf ausgerichtet, auf Basis der OCR-Korrektur und der automatisierten Eigennamenerkennung zusätzlichen Mehrwert für die Benutzerinnen und Benutzer der Plattform E-Periodica zu schaffen. Am einfachsten ist diesbezüglich die Weiterverwendung der verbesserten OCR-Ergebnisse. Hier genügt es, die bestehenden OCR-Textfiles gegen die Files mit den vorgenommenen Korrekturen auszutauschen. Nach einem Update des Index der Plattform steht die optimierte OCR-Erkennung sofort zur Verfügung und trägt zu vollständigeren Trefferlisten bei Suchabfragen bei.

Die Frage, inwiefern die Erkennung von Personen, Ländern und Orten für die Entwicklung neuer Zusatzfunktionen genutzt werden kann, soll unter direktem Einbezug von Usern analysiert werden.

Anhand verschiedener Szenarien ist zu klären, welche Zusatzinformationen von Benutzenden gewünscht werden und welche Verbindungen innerhalb der Plattform und zu externen Informationsquellen von grösstem Nutzen sind. Sobald eine erste Runde der oben erwähnten Tests der Ausweitung des Systems auf zusätzliche Zeitschriften abgeschlossen ist, wird zu diesem Zweck ein Workshop mit Nutzerinnen und Nutzern stattfinden, mit denen die ETH-Bibliothek aufgrund von Anfragen und früheren Rückmeldungen zur Usability der Plattform E-Periodica bereits in Kontakt steht.

Einige Einzelgespräche, die schon mit Usern geführt worden sind, zeigen, in welche Richtung die Entwicklung entsprechender Zusatzfunktionen gehen könnte. So sind etwa die mit hoher Präzision und Ausbeute erkannten Ländernamen und Schweizer Ortsnamen eine sehr gute Ausgangslage, um über einen virtuellen Globus oder eine Landkarte einen geografischen Zugang zu Zeitschriftenartikeln zur Verfügung zu stellen. Auf besonderes Interesse scheinen auch die GND-basierten Verlinkungen von Personen zu stossen. Hier ist denkbar, die GND-ID zu nutzen, um dynamisch sowohl weitere Artikel zu einer Person innerhalb von E-Periodica als auch weiterführende Ressourcen der ETH-Bibliothek (Publikationen, Archivmaterial aus eigenen Beständen) und externe Zusatzinformationen anzuzeigen (vgl. Abb. 2).

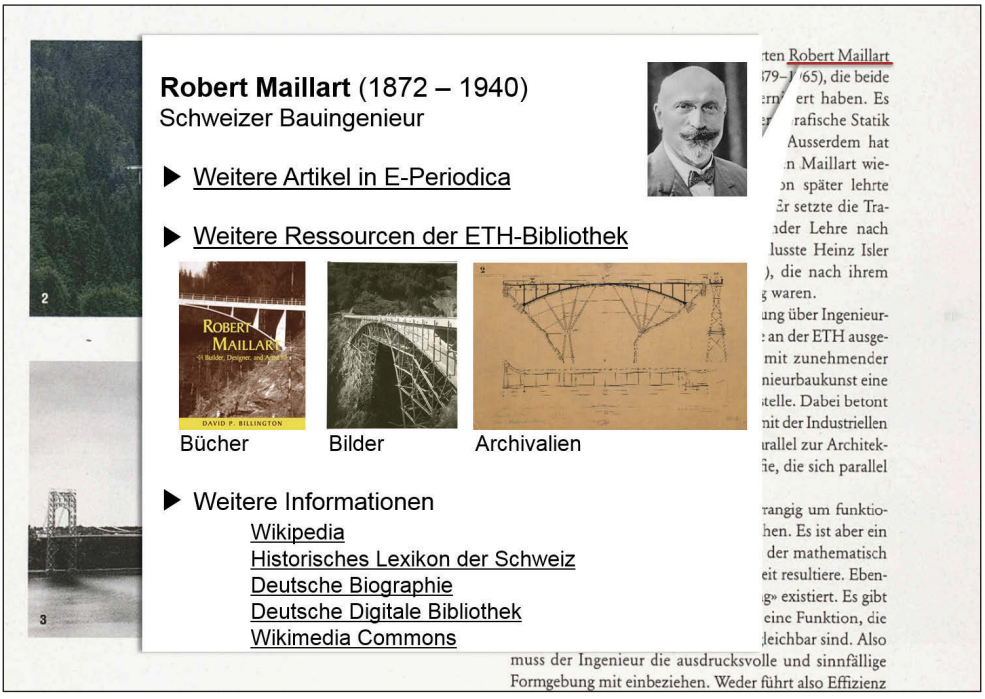

Abb. 2: Mockup einer Anreicherung zu einer Person, die in einem Zeitschriftenartikel automatisch erkannt und mit der GND-ID verlinkt wurde.

In welche Richtung die Entwicklung von Zusatzfunktionen auch gehen wird, sie soll in jedem Fall schrittweise erfolgen und nicht erst dann, wenn die automatische Erkennung über das ganze E-Periodica-Angebot durchgeführt worden ist. Entsprechend deutlich muss für die Benutzenden erkennbar sein, dass es sich um prototypische Features und Anreicherungen handelt, die nur für gewisse Inhalte und keineswegs für alle Personen in sämtlichen Artikeln zur Verfügung stehen.

6. Weiterführende Untersuchungen

Die Diskussion mit Benutzerinnen und Benutzern bezüglich Zusatzfunktionen findet mit Blick auf eine Nachnutzung der Ergebnisse des Pilotprojekts auf einer ganz konkreten und anwendungsorientierten Ebene statt. Den Usern von E-Periodica soll ein direkter Mehrwert geboten werden. Aus der Perspektive des ICL UZH und dessen computerlinguistischer Analyse von E-Periodica als Textkorpus wurden am Ende des Pilotprojekts Fragen formuliert, die auf weiterführende Untersuchungen zu Systemoptimierungen fokussieren. Dazu gehört die Erkennung zusätzlicher Entitäten. Einerseits könnte es interessant sein, im Anschluss an Personen und Orte auch geografische Namen wie Berge, Täler oder Seen zu annotieren. Andererseits wäre auch die Erkennung von Organisationen wie Firmen, Behörden oder Bildungseinrichtungen aufschlussreich. In einem nächsten Schritt könnte die automatische Erkennung von Beziehungen zwischen Personen oder Personen und Orten ein Ziel sein.

Grosses Potential liegt auch in der Mehrsprachigkeit der Inhalte von E-Periodica. So könnten Inhalte, die in mehreren Sprachen vorliegen, etwa genutzt werden, um Ambiguitäten deutscher Namen aufzulösen. Unter Einbezug französischer oder italienischer Übersetzungen liesse sich z. B. die Mehrdeutigkeit von „Zug“ – ist das Verkehrsmittel oder die gleichnamige Stadt in der Innerschweiz gemeint? – klären. Für künftige Weiterentwicklungen von E-Periodica im Austausch mit der computerlinguistischen Forschung bleibt also viel Raum.

7. Fazit

Die Plattform E-Periodica mit ihren über sieben Millionen Seiten an OCR-erkanntem Volltext bietet ein optimales Experimentierfeld im Bereich der automatisierten Eigennamenerkennung. Das vom ICL UZH durchgeführte Pilotprojekt hat gezeigt, dass mit bestehenden computerlinguistischen Ansätzen und Systemen mit vertretbarem Aufwand sowohl OCR-Resultate verbessert als auch zentrale Entitäten wie Personen, Länder- und Ortsnamen erkannt und zu einem gewissen Teil verlinkt werden können. Auf dieser Basis baut die ETH-Bibliothek ihre internen Kompetenzen in den Bereichen *OCR-Postprocessing*, *Named Entity Recognition* und *Named Entity Linking* gezielt aus. Im Vordergrund steht dabei, den Benutzenden der Plattform E-Periodica die angereicherten Daten und erkannten Eigennamen möglichst gewinnbringend anzubieten. In welcher Form und mit welchen Zusatzfunktionalitäten das geschieht, wird im direkten Austausch mit Usern ermittelt. Darüber hinaus bleibt aber das beachtliche und laufend wachsende E-Periodica-Textkorpus auch für weitergehende computerlinguistische Forschung interessant.

Literaturverzeichnis

- von Däniken, Pius; Cieliebak, Mark: Transfer Learning and Sentence Level Features for Named Entity Recognition on Tweets, in: The Association for Computational Linguistics (Hg.), Proceedings of the 3rd Workshop on Noisy User-generated Text, Copenhagen, Denmark, September 7, 2017, S. 166–171. Online: <http://www.aclweb.org/anthology/W17-4422>, Stand: 24.09.2018.

- Ebling, S; Sennrich, R; Klaper, D; Volk, Martin: Digging for names in the mountains: Combined person name recognition and reference resolution for German alpine texts, in: 5th Language & Technology Conference, Poznan, Poland, 25 November 2011 - 27 November 2011. Online: <https://doi.org/10.5167/uzh-50451>.

- ETH-Bibliothek Zürich (Hg.): ETH-Bibliothek Jahresbericht 2016, Zürich 2017. Online: <https://doi.org/10.3929/ethz-a-004157606>.

- Schmid, Helmut: TreeTagger – a part-of-speech tagger for many languages, <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>, Stand: 24.09.2018.

