Statistische Analysen von Semantic Entities aus Metadaten- und Volltextbeständen von German Medical Science

Statistical evaluation of semantic entities from metadata and full texts on German Medical Science corpora

Stefan Grün[1]
Christoph Poley[1]

1 ZB MED – Informationszentrum Lebenswissenschaften, LIVIVO Entwicklung, Köln, Deutschland

Abstract

This paper analyzes the information content of metadata and full texts in German Medical Science (GMS) articles in English language. The object of the study is to compare semantic entities that are used to enrich GMS metadata (titles and abstracts) and GMS full texts.
The aim of the study is to test whether using full texts increases the value added information. The comparison and evaluation of semantic entities was done statistically. Measures of descriptive statistics were gathered for this purpose. In addition to the ratio of central tendencies and scatterings, we computed the overlaps and complements of the values.
The results show a distinct increase of information when full texts are added. On average, metadata contain 25 different entities and full texts 215. 89% of the concepts in the metadata are also represented in the full texts. Hence, 11% of the metadata concepts are found in the metadata only. In summary, the results show that the addition of full texts increases the informational value, e.g. for information retrieval processes.

Keywords: ZB MED, LIVIVO, ZB MED Knowledge Environment, ZB MED KE, metadata, full texts, named entitiy recognition, UIMA, SPSS, descriptive statistics, MongoDB, life science

Zusammenfassung

Dieser Fachbeitrag beschäftigt sich damit, englischsprachige German Medical Science (GMS) Artikel zu analysieren. Untersuchungsgegenstand ist ein Vergleich zwischen Semantic Entities, mit denen GMS-Metadaten (Titel und Zusammenfassungen) und GMS-Volltexte angereichert werden. Inwieweit der informationelle Mehrwert durch Hinzunahme von Volltexten steigt, ist Fragestellung dieses Beitrages. Der durchgeführte Vergleich erfolgt statistisch durch die Auswertung annotierter Semantic Entities. Es werden hierzu Kennziffern der deskriptiven Statistik berechnet. Neben den Kennziffern zur zentralen Tendenz und zur Streuung erfolgt zudem eine Berechnung der Schnitt- und Differenzmengen.
Die Ergebnisse zeigen ein deutliches Mehr an Informationen aus den Volltexten. Durchschnittlich liegen in den Metadaten 25 verschiedene Entities vor, in den Volltexten hingegen 215. 89% der Konzepte aus den Metadaten werden auch im Volltext repräsentiert. Dagegen werden 11% der gefundenen Konzepte der Metadaten auch nur in den Metadaten gefunden. Die berechneten Ergebnisse belegen statistisch, dass durch die Hinzunahme von Volltexten der informationelle Mehrwert z.B. für das Information Retrieval steigt.

Schlüsselwörter: ZB MED, LIVIVO, ZB MED Knowledge Environment, ZB MED KE, Metadaten, Volltexte, Eigennamenerkennung, UIMA, SPSS, deskriptive Statistik, MongoDB, Lebenswissenschaften

Einleitung

Die Deutsche Zentralbibliothek für Medizin (ZB MED) – Informationszentrum Lebenswissenschaften (https://www.zbmed.de/) versteht sich als zentrale Informationsinfrastruktur für die Lebenswissenschaften in Deutschland und Europa. Aufbauend auf einzigartige Bestände bietet ZB MED forschungsbasierte Möglichkeiten zur Gewinnung von Informationen und Nutzung von Forschungsdaten in den Lebenswissenschaften. Für die Recherche nach wissenschaftlicher Fachliteratur wird das Suchportal LIVIVO entwickelt und betrieben (https://www.livivo.de), das den Zugriff auf etwa 60 Millionen Medien aus dem Bereichen Medizin- und Gesundheitswesen sowie den Umwelt-, Ernährungs- und Agrarwissenschaften bietet. Als Datengrundlage dient das ZB MED Knowledge Environment (ZB MED KE) [1], in dem mehr als 70 verschiedene Datenquellen aus den Lebenswissenschaften zusammengeführt werden. Dieses stellt u.a. die Datengrundlage für LIVIVO bereit und beinhaltet neben den dort abrufbaren Metadaten beispielsweise auch Volltexte, Verfügbarkeiten oder Verknüpfungen zu aktuellen Forschungsarbeiten [2].

Für die vorliegende Untersuchung sind Forschungsarbeiten zur Anreicherung des ZB MED Datenbestands mit sogenannten Semantic Entities (Begriffe, Konzepte, Deskriptoren oder vereinfacht Entities [3]) grundlegend. Diese sind Ein- oder Mehrwortbenennungen aus einem kontrollierten Vokabular. Sie beinhalten allgemeinsprachliche Sachbegriffe (z.B. aus den Thesauri MeSH [4] oder Agrovoc [5]) oder Individualnamen (z.B. in DrugBank [6]). In beiden Fällen verfügen sie über die Charakteristika, kontextfrei, wiedergabetreu und vorhersagbar zu sein [7]. Deshalb können sie auch im Information Retrieval als Indexterme eingesetzt werden.

Eine Studie von Müller et al. [8] beschäftigt sich damit, auf der Basis von Titeln und Zusammenfassungen Semantic Entities zu generieren und den Dokumenten zuzuordnen. Zudem wurde ein Wordcloud-Prototyp (http://labs.livivo.de) vorgestellt, der die häufigsten Entities je Wörterbuch im LIVIVO-Korpus visualisiert. Eine zweite Studie von Müller et al. [9] befasst sich damit, die identifizierten Semantic Entities zu zählen und dabei den jeweiligen Anteil der Wörterbücher mithilfe des Jaccard-Index zu messen. Ergebnis hier war beispielsweise eine hohe Schnittmenge bei den allgemeinen Sachdeskriptoren der Thesauri MeSH und Agrovoc, während die Datenbank DrugBank aufgrund ihrer Individualnamen niedrige Indexzahlen zur Ähnlichkeit der identifizierten Konzepte aufwies.

Die vorliegende Studie knüpft an die oben genannten Arbeiten an und weitet in einem ersten Schritt den Ansatz zum Generieren von Semantic Entities für Metadaten [8] zusätzlich auf Volltexte aus. Verwendet werden hierbei englischsprachige Fachartikel aus German Medical Science. Ein statistischer Vergleich zwischen Semantic Entities aus Metadaten und Volltexten erfolgt dann im zweiten Schritt der Arbeit. Hierzu werden neben Kennwerten der deskriptiven Statistik zusätzlich Schnitt- und Differenzmengen berechnet. Hintergrund ist dabei die Fragestellung, inwieweit eine Hinzunahme des Volltextes einen informationellen Mehrwert darstellt, beispielsweise für das Information Retrieval lebenswissenschaftlicher Literatur in LIVIVO.

Methoden

Das ZB MED KE beinhaltet die technische Infrastruktur zum Datenmanagement und dient damit als Ausgangspunkt für die vorliegenden Untersuchung. Dessen Aufgabe ist es unter anderem, Daten aus verschiedenen Quellen (z.B. ZB MED-Kataloge, Medline [10], DissOnline [11], German Medical Science [12] oder Verlagsdaten) und unterschiedlichen Datenformaten (z.B. MARC21 [13], JATS [14] oder DublinCore [15]) einzusammeln, zu harmonisieren und für weitere Anreicherungen abzuspeichern. Die Daten dienen auch als Grundlage für die anwendungsorientierte Forschung an ZB MED und/oder Dienstleistungen wie LIVIVO. Das dokumentenbasierte Datenbankmanagementsystem MongoDB (http://www.mongodb.com) bildet hierbei die Infrastruktur für die Datenhaltung des ZB MED KE [2]. Über einen Connector erfolgt der Zugriff auf die darin enthaltenen GMS-Dokumente (Metadaten und Volltexte), die als Grundlage für die Studie dienen.

Die im ZB MED KE vorliegende Gesamtzahl der GMS-Artikel wurde zunächst auf 530 Dokumente eingeschränkt. Die Filterung erfolgte auf die Artikeltypen „Case Report“, „Research Article“ und „Review Article“. Poster, Abstracts ohne Volltexte oder ähnliche Formate wurden nicht herangezogen. Darüber hinaus beschränkte sich die Studie auf englischsprachige Dokumente, da die eingebundenen Wörterbücher (MeSH, Agrovoc und DrugBank) ebenfalls in englischer Sprache zum Einsatz kamen.

Im zweiten Schritt wurden die im ZB MED KE abgelegten Dokumente extrahiert und verarbeitet. Ausgangspunkt hier waren Dokumente, die sich verallgemeinert in die Kategorien Metadaten und Volltexte unterteilen. In der Kategorie Metadaten finden sich inhaltliche (z.B. Titel oder Abstract) und formale (z.B. DOI oder Autor) Aspekte. Die Kategorie Volltexte beinhaltet XML-Repräsentationen der ausgewählten GMS-Dokumente. Um einen Vergleich zwischen Semantic Entities aus Metadaten (Titel und Zusammenfassung) und Volltexten zu realisieren, wurden die XML-Auszeichnungen entfernt.

Im nächsten Schritt ließ sich der auf das UIMA-Framework (https://uima.apache.org) basierende Concept Mapper (vgl. vorausgegangene Studie von Müller et al. [8]) einsetzen und damit die Metadaten (Titel und Zusammenfassung) sowie die Volltexte annotieren. Die eingebundenen Wörterbücher waren MeSH, Agrovoc und DrugBank. Das MeSH-Vokabular besteht aus allgemeinen Sachdeskriptoren der Medizin, der multilinguale Agrovoc Thesaurus aus allgemeinen Sachbegriffen der Agrarwirtschaft. DrugBank beinhaltet Individualnamen von Medikamenten. Alle drei Ressourcen realisieren eine Synonymrelationierung. Dementsprechend konnte die Zuteilung der Deskrip-

toren (Vorzugsbenennungen) zu Dokumenten durch das UIMA Framework durchgeführt werden. Im MeSH-Thesaurus wurden beispielhaft u.a. die bedeutungsgleichen Benennungen „Tumors“ und „Cancer“ der Vorzugsbenennung „Neoplasms“ zugeordnet. Taucht in einem Dokument eines dieser Synonyme oder die Vorzugsbenennung selbst auf, wird der Begriff „Neoplasms“ identifiziert und dem jeweiligen Dokument zugeordnet. Die Ergebnisse des UIMA Concept Mappers wurden im ZB MED KE für die weitere Verarbeitung zusätzlich abgelegt [2].

Nach Anreicherung der Metadaten und Volltexte durch Semantic Entites erfolgte die Extraktion (getrennt nach Metadaten und Volltexten) und die Zählung der annotierten Begriffe. Für den vorliegenden Vergleich wurden alle Begriffe beim ersten Vorkommen einfach gezählt, da die Untersuchung einen möglichen informationellen Mehrwert anhand unterschiedlicher Entities analysiert. Zudem erfolgte an dieser Stelle die Messung von Schnitt- und Differenzmengen (vgl. Tab. 2). Wenn ein identifizierter Begriff im Titel oder in der Zusammenfassung (Metadaten) auftaucht und zusätzlich auch im Volltext steht, wird dieser als Schnittmenge eingetragen. Sofern der Begriff nur in den Metadaten oder nur im Volltext steht, wird dieser als Komplement eingeordnet. In dem am Ende erzeugten Extraktionsergebnis befanden sich somit Semantic Entities aus Metadaten und aus Volltexten sowie deren Schnitt- und Differenzmengen. Jeweils hinter den Begriffen stand deren errechnete Anzahl der unterschiedlichen Entities je Dokument.

Die erzeugte Datei wurde dann in ein SPSS konformes Format gebracht und statistisch analysiert (https://www.ibm.com/analytics/de/de/technology/spss). Hier folgten Berechnungen der deskriptiven Statistik [16], um die vorliegenden Begriffsmengen hinsichtlich der im ersten Abschnitt formulierten Forschungsfrage zu untersuchen. Die Kennwerte der zentralen Tendenz (Modus, Median und arithmetisches Mittel) waren relevant, um zu analysieren wie viele unterschiedliche Konzepte je Dokument im Mittel annotiert wurden. Die Werte der Verteilung/Streuung (Standardabweichung, Varianz, Spannweite, Minimum und Maximum sowie Summe) sollten zeigen, wie repräsentativ die Werte der Mitte waren und wie diese, bezogen auf die gesamte Stichprobe von 530 Dokumenten einzuordnen sind.

Ergebnisse

Tabelle 1 fasst die durch SPSS berechneten Ergebnisse für die Konzepte der Metadaten und der Volltexte zusammen.

Am häufigsten (Modus) wurden zu den Metadaten 22 Begriffe (vgl. Abbildung 1) annotiert. Bei den Volltexten wird – mit Verweis auf eine gleich häufige Anzahl von Begriffen – ein Wert von 98 Begriffen (vgl. Abbildung 2) angegeben. 98 Begriffe wurden insgesamt sechsmal identifiziert – daneben tauchen 132, 136, 137, 146, 172, 181 und 205 ebenfalls sechsmal auf. Das Verhältnis der Modi zwischen Metadaten (22) und Volltexten (98) liegt bei 4,5. Mit dem größten Modus für die Volltexte (205) wird ein Verhältnis von 9,3 erreicht.

Tabelle 1: Werte der deskriptiven Statistik zu den identifizierten Semantic Entities aus Metadaten- und Volltextbeständen des GMS-Korpus (N=530)

	Metadaten	Volltexte
arithmetisches Mittel	25,23	214,83
Median	24,00	175,00
Modus	22,00	98,00*
Standardabweichung	10,86	141,59
Varianz	117,85	20.048,05
Spannweite	80,00	1.104,00
Minimum	5,00	37,00
Maximum	85,00	1.141,00
Summe	13.374,00	113.862,00

* Mehrere Modi vorhanden. Angabe des kleinsten Werts

Der Zentralwert (Median) liegt in den Metadaten bei 24 und in den Volltexten bei 175 Entities. Für die Metadaten bedeutet dies, das die untere Hälfte der 530 Dokumente gleich viele oder weniger als 24 Entities und die obere Hälfte der Dokumente gleich viele oder mehr als 24 Entities identifiziert wurden. Analog dazu beinhalten die unteren 50% der 530 Volltexte gleich viele oder weniger als 175 Entities und die oberen 50% der Dokumente gleich viele oder mehr als 175 Entities. Hier liegt der Faktor der Mediane von Volltexten zu Metadaten bei ca. 7,3.

Das arithmetische Mittel bei den Metadaten beträgt 25,23 und 214,83 bei den Volltexten. Es wurden damit auf die verwendeten 530 Dokumente durchschnittlich 25 Begriffe in den Metadaten und 215 Begriffe in den Volltexten identifiziert. Die Anzahl an gefundenen Konzepte aus den Volltexten ist mit einem Faktor von ca. 8,6 höher als die Anzahl aus den Metadaten.

Die Standardabweichung 10,86 (Varianz: 117,85) in den Metadaten zeigt, dass die Häufigkeiten der identifizierten Sematic Entities in den Bereichen zwischen 14,37 (Mittelwert – Standardabweichung) und 36,09 (Mittelwert + Standardabweichung) liegen. Für die Volltexte wurde eine Standardabweichung von 141,59 (Varianz: 20.048,05) Konzepten ermittelt. Hier liegen die meisten Häufigkeiten der erkannten Semantic Entities zwischen 73,24 und 366,42.

In den Metadaten lag in einer Range (Spannweite) von 80 Konzepten ein Minimum von 5 und ein Maximum von 85 Konzepten je Dokument vor. Für die Volltexte wurden in einer Spannweite von 1.104 Konzepten ein Minimum von 37 und ein Maximum von 1.141 Konzepten berechnet. Hier liegen die Werte des Minimums zwischen Metadaten (5) und Volltexten (37) ca. um den Faktor 7 auseinander, während die Werte des Maximums zwischen Metadaten (85) und Volltexten (1.141) ca. den Faktor 13 betragen.

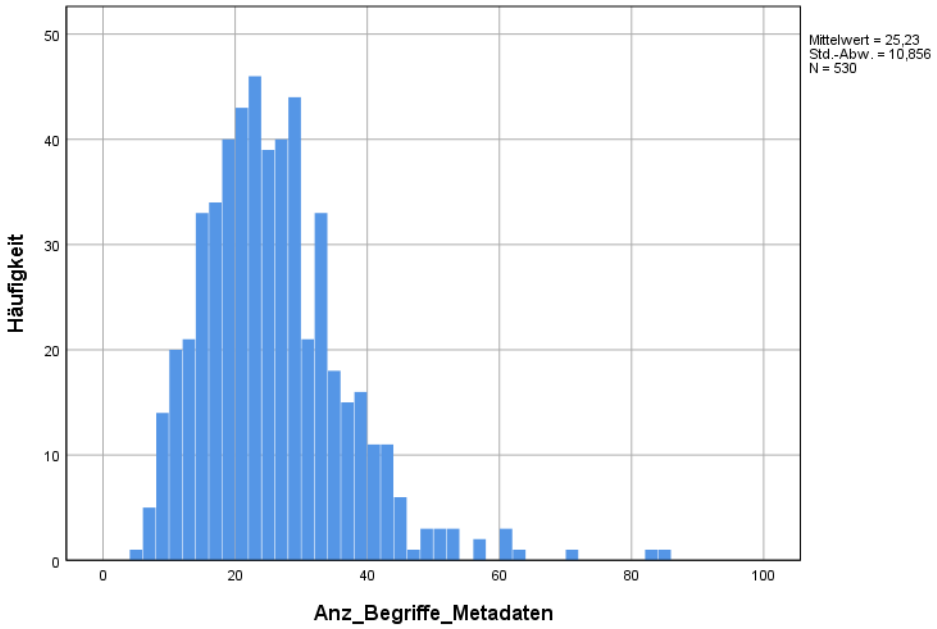

Abbildung 1: Histogramm zur Anzahl der identifizierten Begriffen in den Metadaten der GMS-Dokumente

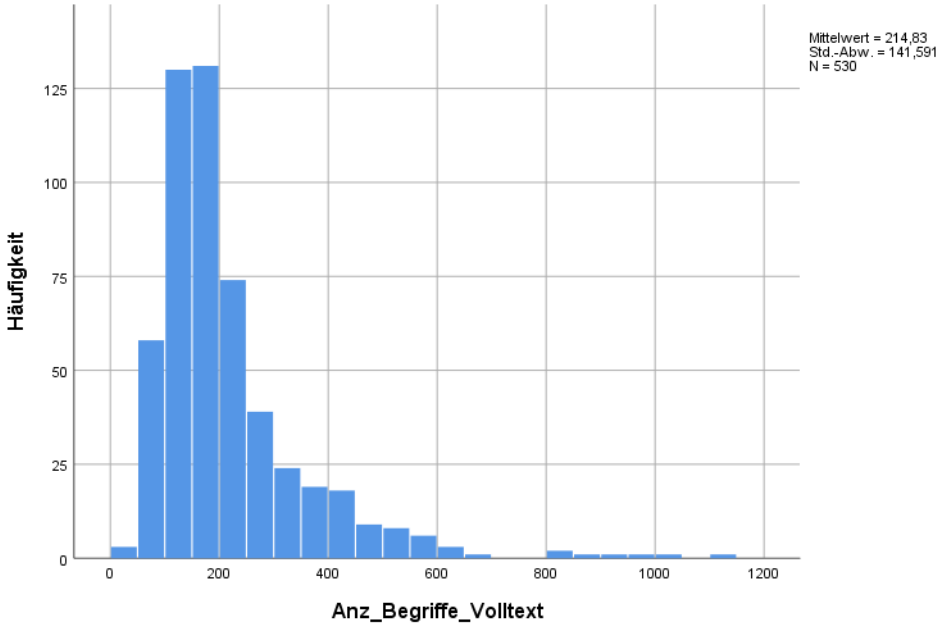

Abbildung 2: Histogramm zur Anzahl der identifizierten Begriffe in den Volltexten der GMS-Dokumente

Die unterschiedlichen Kennwerte der zentralen Tendenz (vgl. Tabelle 1) zeigen für die Metadaten geringe Abweichung untereinander – der Median (24) weicht um zwei Konzepte vom Modus (22) und um 1,23 Konzepte vom arithmetischen Mittel (25,23) ab. Deutlicher sind die Abweichungen in den Werten der Volltexte. Der arithmetische Mitteilwert (214,83) weicht um 39,80 vom Median ab und liegt 116,83 vom Modus entfernt, wobei hier zu beachten ist, das 98 der niedrigste angegebene Wert ist. Die Kennzahlen der Streuung zeigen, dass vor allem in Volltexten Dokumente mit hoher Anzahl an identifizierten Konzepten gefunden wurden. Dies zeigt z.B. die Standardabweichung der Volltexte (141,59) vom Mittelwert (214,83), bei einem höheren Maximum (1.141). Bei den Metadaten zeigt die Standardabweichung (10,86) vom Mittelwert (25,23) bei dem größeren Maximum (85) ebenfalls vereinzelt Datensätze mit vielen identifizierten Konzepten.

Tabelle 2 zeigt das Verhältnis der in dem GMS-Testkorpus identifizierten Entities. In den Metadaten der 530 GMS-Dokumente befinden sich insgesamt 13.374 gefundene Konzepte. Die Volltexte weisen eine Summe von 113.862 Konzepten auf. Das stellt einen Faktor von ca. 8,5 dar.

Von den in den Metadaten gefundenen 13.374 Begriffen befinden sich 11.902 (ca. 89%) in den Volltexten. Die verbleibenden 1.472 Begriffe (ca. 11%) der Metadaten wurden nicht in den Volltexten gefunden. Analog dazu beinhalten die Volltexte – mit 101.960 von insgesamt 113.862 Begriffen (ca. 89%) – Begriffe, die nicht in den Metadaten zu finden sind.

Tabelle 2: Schnitt- und Differenzmenge der Semantic Entities aus Metadaten- und Volltextbeständen des GMS-Korpus (N=530)

	Metadaten ∩ Volltexte	Metadaten / Volltexte	Volltexte / Metadaten
Summe	11.902	1.472	101.960

Diskussion

Die statistischen Berechnungen zeigen, dass nach dem angewendeten Verfahren auf der Basis von identifizierten Konzepten in Volltexten ein Mehr an Informationen als in Metadaten vorhanden ist. Dies lässt sich anhand der höheren Werte aus den vorliegenden statistischen Berechnungen durchgängig belegen. Beispielsweise zeigt der Median für die Volltexte mit 175 Konzepten im Vergleich zum Median der Metadaten mit 24 Konzepten einen um den Faktor 7 höheren Wert. Ergänzend dazu bestätigt der Vergleich der Maximalwerte von Metadaten (85 Begriffe) und Volltexten (1.141 Begriffe) die Annahme.
Eine mögliche Erklärung dafür ist, dass Metadaten (hier Titel und Zusammenfassung) nur ein Kurzabriss über den gesamten Text darstellen. Verfasser beschränken sich in der Regel in Metadaten auf das Grundlegendste und führen oft erst im Verlauf des (Voll-)Textes alle Fachbegriffe ein, beschreiben ihre Methoden und analysieren und reflektieren ihre Erkenntnisse ausführlich.
Weiterhin zeigen die Ergebnisse, dass ein Großteil der in den Metadaten befindlichen Informationen auch in den Volltexten repräsentiert ist. 89% der in den Metadaten identifizierten Begriffe sind auch im jeweiligen Volltext gefunden worden. Dies deutet darauf hin, dass die wohl grundlegenden (Fach-)Begriffe in den Metadaten untergebracht werden. Dies liegt im Interesse eines Verfassers, da Titel und Zusammenfassung vorzugsweise vor dem eigentlichen Volltext von möglichen interessierten Lesern durchgesehen werden. Außerdem richtet sich ein Fachbeitrag in der Regel an eine Fachleserschaft, bei der bestimmte Fachbegriffe nicht erklärt, sondern nur in einen bestimmten Zusammenhang gebracht werden müssen.
Trotz der großen Schnittmenge beinhalten Metadaten Informationen, die nicht in Volltexten repräsentiert werden. Die Kennziffer der Differenzmengen aus Tabelle 2 zeigt, dass von den 13.374 Begriffen 1.472 (ca. 11%) nicht im Volltext benannt werden. Dies kann etwa dafür sprechen, dass bestimmte Individualnamen nur in Metadaten genannt werden oder dass es sich dabei um allgemeine Sachdeskriptoren handelt, die im Volltext nicht vorkommen.
Eine nicht zu vernachlässigende Bedingung für die konkreten Werte aus den statistischen Berechnungen besteht in der Charakteristik der Entities in den Volltexten und Metadaten, die im Rahmen dieser Arbeit aber kein Untersuchungsgegenstand war. Eine angelehnte Analyse unter gleichen Bedingungen mit der Untersuchung getrennt nach wörterbuchspezifischen Konzepten könnte diese Annahme bestätigen. Ein mögliches Resultat kann dann sein, dass die Verwendung von allgemeinen Namen und Individualnamen in den Metadaten und den Volltexten unterschiedlich ist.
Werden mehrere Wörterbücher gleichzeitig eingesetzt, dann wird dieses Verhältnis die Ergebnisse auch beeinflussen. Müller et al. [9] zeigten beispielsweise in ihrer Arbeit, dass unter den Top1000 Deskriptoren die Schnittmengen zwischen den beiden allgemeinsprachlichen Thesauri MeSH und AGROVOC (473 von 1.000 Konzepten) deutlich höher sind, als z.B. zwischen einem allgemeinsprachlichen Thesaurus MeSH und den Individualnamen in der DrugBank (21 von 1.000 Konzepten). Übertragen auf diese Studie kann der Einsatz von zum Beispiel zwei allgemeinsprachlichen Thesauri (MeSH und AGROVOC) jeweils eine quantitative Auswirkung auf das Identifizieren von Deskriptoren haben.
Ergänzend dazu sei ein weiterer statistischer Aspekt eingefügt: Im Titel wurde im Schnitt alle 24 Zeichen ein neuer Begriff identifiziert, in der Zusammenfassung tauchte ein neuer Begriff alle 35 Zeichen und im Volltext alle 43 Zeichen auf. Den Metadaten standen im Durchschnitt mit 33 Zeichen die Volltexte mit 43 Zeichen gegenüber. In Metadaten herrscht wie erwartet ein etwas höherer Wert an identifizierten Entities als in Volltexten bezogen auf die Anzahl von Zeichen.

Schlussfolgerungen

Der vorliegende Fachbeitrag belegt anhand der verwendeten Methoden auf Basis von Semantic Entities statistisch, dass Volltexte einen informationellen Mehrwert gegenüber Metadaten besitzen. Trotz einiger Einschränkungen dieser Studie, z.B. die Begrenzung auf eine Kollektion anstatt einer Ausweitung auf mehrere Quellen, die Anzahl der Datensätze (530 Dokumente) oder die Beschränkung auf 3 Wörterbücher, ist eine deutliche positive Tendenz für die Hinzunahme des Volltextes erkennbar. Das genutzte Verfahren bietet damit einen zusätzlichen, möglicherweise auch einen alternativen Aspekt zu Diskussionen um Metadaten- und/oder Volltextindexierungen (z.B. [17], [18]).
Dabei bietet die Anreicherung von Metadaten und Volltexten durch Semantic Entities in ihrer Gesamtheit einen Mehrwert. Individualnamen sind spezifischer als Allgemeinwörter und damit potenziell stärker für ein Information Retrieval geeignet – besonders vor dem Hintergrund, dass LIVIVO sich an eine lebenswissenschaftliche Community richtet. Bei allgemeinen Sachdeskriptoren könnte u.U. auf eine Anreicherung verzichtet werden, sofern die Allgemeinwörter trotz semantischer Disambiguierung zu allgemein für ein ballastfreies Information Retrieval sind. Auch hier könnte eine anknüpfende wörterbuchspezifische Analyse der Deskriptoren erkenntnisreich sein.
Ein Ansatzpunkt für ZB MED wäre, Autoren von Fachbeiträgen bei der Wahl des Titels sowie dem Erstellen der Zusammenfassung (Abstract) in Zukunft durch automati-

sche Verfahren zu unterstützen. Dies kann zum Beispiel realisiert werden, indem der vom Autor eingereichte Aufsatz dahingehend analysiert wird, dass ein Verhältnis der gefundenen Konzepte in seinen Metadaten und dem Volltext gebildet wird, mit dem Ziel, einen bestimmten Wert nicht zu überschreiben. Als Referenzwert könnte hier exemplarisch das Verhältnis von 1/7 für die Anzahl der gefundenen Entities in den Metadaten/Volltexte dienen. Damit kann der Autor statistisch eine „bessere" Berücksichtigung seines Aufsatzes beim Information Retrieval auf Metadaten erzielen. Dies kann dazu beitragen, dass wissenschaftliche Aufsätze möglichst wiedergabetreu beschrieben werden. Darüber hinaus ließe sich ein weiteres geeignetes Werkzeug einsetzen, mit dem Ziel, automatisch ergänzende Entities für die Metadaten vorzuschlagen. Dies ließe sich dann zum Beispiel durch ein Highlighting der relevanten Textpassagen in den Volltexten mit Verlinkung auf die entsprechenden dahinter stehenden Entities realisieren.

Anmerkung

Interessenkonflikte

Die Autoren arbeiten an ZB MED in den Informationsdiensten für die LIVIVO-Entwicklung. Herr Poley ist darüber hinaus für das Produkt LIVIVO verantwortlich. Ansonsten liegen keine Interessenkonflikte vor.

Literatur

1. Müller B, Poley C, Pössel J, Hagelstein A, Gübitz T. LIVIVO – the Vertical Search Engine for Life Sciences. Datenbank-Spektrum. 2017;17:29-34.
2. Poley C. Neue Herausforderungen an das ZB MED-Suchportal für Lebenswissenschaften. GMS Med Bibl Inf. 2016;16(3):Doc21. DOI: 10.3205/mbi000376
3. Stock WG. Concepts and semantic relations in information science. J Assoc Inf Sci Technol. 2010;61(10):1951-69. DOI: 10.1002/asi.21382
4. National Library of Medicine (NLM). Medical Subject Headings. [letzter Zugriff: 15.12.2017]. Verfügbar unter: https://www.ncbi.nlm.nih.gov/mesh
5. Food and Agriculture Organization (FAO) of the United Nations. AGROVOC Multilingual agricultural thesaurus. [letzter Zugriff: 15.12.2017]. Verfügbar unter: http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus
6. University of Alberta. DrugBank. [letzter Zugriff: 15.12.2017]. Verfügbar unter: https://www.drugbank.ca/
7. Gödert W, Lepsky K, Nagelschmidt M. Informationserschließung und automatisches Indexieren – ein Lehr- und Arbeitsbuch. Berlin, Heidelberg: Springer; 2012.
8. Müller B, Hagelstein A. Beyond Metadata – Enriching Life Science Publications in LIVIVO with Semantic Entities from the Linked Data Cloud. In: SEMANTicS; 2016 Sep 12-15; Leipzig, Germany.
9. Müller B, Hagelstein A, Gübitz T. Life science ontologies in literature retrieval: A comparison of linked data sets for use on semantic search on a heterogeneous corpus. In: Proceedings of the 20th International Conference on Knowledge Engineering and Knowledge Management; 2016 Nov 19-23; Bologna, Italy.
10. National Library of Medicine (NLM). Medline. [letzter Zugriff: 15.12.2017]. Verfügbar unter: https://www.ncbi.nlm.nih.gov/pubmed/
11. Deutsche National Bibliothek (DNB). Dissonline. [letzter Zugriff: 15.12.2017]. Verfügbar unter: http://www.dnb.de/DE/Wir/Kooperation/dissonline/dissonline_node.html
12. Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF). German Medical Science GMS. [letzter Zugriff: 15.12.2017]. Verfügbar unter: http://www.egms.de/
13. Deutsche National Bibliothek (DNB). MARC 21. [letzter Zugriff: 15.12.2017]. Verfügbar unter: http://www.dnb.de/DE/Standardisierung/Formate/MARC21/marc21_node.html
14. JATS. Journal Article Tag Suite. [letzter Zugriff: 15.12.2017]. Verfügbar unter: https://jats.nlm.nih.gov/
15. Dublin Core Metadata Initiative (DCMI). Dublin Core. [letzter Zugriff: 15.12.2017]. Verfügbar unter: http://dublincore.org/
16. Field A. Discovering statistics using SPSS. 3. Aufl. London: Sage Publications; 2009.
17. Stock WG. Automatische und intellektuelle Indexierung. Das „Big-Systems'-Syndrom". Eine Antwort auf Robert Fugmann. Password. 2000;(1):18-20.
18. Beall J. The weaknesses of full-text searching. J Acad Librariansh. 2008;34(5):438-44.

Erratum

In der Titelübersetzung wurde ein Buchstabendreher (semanitc) korrigiert.

Korrespondenzadresse:
Christoph Poley
ZB MED – Informationszentrum Lebenswissenschaften, Leitung LIVIVO, Gleueler Str. 60, 50931 Köln, Deutschland
poley@zbmed.de

Bitte zitieren als
Grün S, Poley C. Statistische Analysen von Semantic Entities aus Metadaten- und Volltextbeständen von German Medical Science. GMS Med Bibl Inf. 2017;17(3):Doc14.
DOI: 10.3205/mbi000393, URN: urn:nbn:de:0183-mbi0003935

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/mbi/2017-17/mbi000393.shtml

Veröffentlicht: *20.12.2017*
Veröffentlicht mit Erratum: *21.12.2017*