Universität Leipzig
Institut für Informatik
Studiengang Data Science, M.Sc.

Erstellung eines umfangreichen Korpus von extrahierten medizinischen Entitäten aus bibliografischen Daten

Masterarbeit

Gregor Pfänder

1. Gutachter: Jun.-Prof. Dr. Martin Potthast
2. Gutachter: Ferdinand Schlatt

Datum der Abgabe: 17. Juni 2023

Zusammenfassung

Die medizinische Informatik bietet viele Anwendungsfälle für Natural Language Processing. Eine wichtige Grundlage, um diese umsetzen zu können, sind Datensätze zum Entwickeln von Modellen. Das Ziel der vorliegenden Arbeit ist es, eine Methodik zur Erstellung eines umfangreichen Korpus vorzustellen. Der Korpus besteht aus einer Auswahl mehrerer Datensätze in denen Textstellen markiert und dazu passende UMLS-Konzepte zugeordnet sind. Dafür wurden mehr als 20 Millionen Abstracts der PubMed-Baseline aus dem Jahr 2021 herangezogen. Um die Datensätze zu erstellen, wurden drei verschiedene Ensemble-Methoden entwickelt, die es ermöglichen, die Ausgaben mehrerer Entity-Linker in einem Ergebnis zusammenzuführen. Die verwendeten Entity-Linker sind scispaCy, QuickUMLS und cTAKES. Zur Evaluation der Ensemble-Methoden wurden drei Evaluationsschritte durchlaufen. Für den ersten Teil wurden verschiedene Kennwerte verglichen. Als Vergleichsdatensatz mit Goldstandard wurde MedMentions gewählt. Da die Rahmenbedingungen bei MedMentions und dem vorgestellten Korpus nicht komplett gleich sind, wurde im Anschluss eine händische Zweitevaluation durchgeführt. Beim dritten Teil der Evaluation wurden, um die Merkmale der Datensätze zu vergleichen, deskriptive Statistiken erstellt. Die Konstruktion des Korpus legt einen Grundstein für viele folgende Projekte und liefert wichtige Ressourcen zur Umsetzung von Anwendungsfällen in der Biomedizin.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Wissensgraphen sind Datenstrukturen, in denen die Entitäten eines Wissensgebiets beschrieben und über Relationen miteinander verbunden werden [17]. Sie können genutzt werden, um die Informationen eines Textes besser zu verstehen. Wissensgraphen liefern Begriffsdefinitionen, Synonyme und zeigen die Beziehungen und Abhängigkeiten von Begriffen untereinander auf. Es gibt viele Anwendungsfälle, die sich Wissensgraphen zu nutzen machen. Zum Beispiel können die Entitäten zum Indizieren von Texten für eine Suchmaschine oder als Unterstützung beim Entwickeln von Sprachmodellen dienen. Einer der relevantesten Wissensgraphen in der Biomedizin ist das Unified Medical Language System (UMLS) [6]. Es vereint das Wissen von mehr als 100 Lexika aus verschiedenen biomedizinischen Domänen und fügt es in einem System zusammen. Das UMLS bildet eine Grundlage, um die angesprochenen Anwendungsfälle für Wissensgraphen auch in der Biomedizin umsetzen zu können. Dafür braucht es aber häufig spezielle Datensätze. Bei einer Familie dieser Datensätze handelt es sich um Texte mit biomedizinischen Inhalten, in denen an passenden Stellen UMLS-Entitäten markiert sind.

Zwei Beispiele solcher Datensätze sind MedMentions und PubMedDS. MedMentions ist ein qualitativ hochwertiger Datensatz mit Goldstandard [25]. Aufgrund hoher Kosten beim händischen Annotieren des Datensatzes ist MedMentions mit 4.392 Dokumenten allerdings nicht besonders groß. Im Gegensatz dazu steht der PubMedDS-Datensatz. Für PubMedDS wurde ein automatisierter Ansatz zum Annotieren der Daten verfolgt [36]. Dadurch war es möglich, einen Datensatz mit ca. 13 Mio. Dokumenten zu erstellen. Das Vorgehen bei PubMedDS führt dabei jedoch zu einer geringen Markierungsdichte. Bei einem direkten Vergleich der beiden Datensätze sind nur 5,3% der in MedMentions markierten Textstellen auch in PubMedDS markiert. Das Ziel der vorliegenden Arbeit ist es, auf automatisierte Weise einen Korpus von mehreren artgleichen

Datensätzen mit extrahierten UMLS-Entitäten zu erstellen. Die Datensätze sollen in ihrer Größe PubMedDS ähneln und gleichzeitig bei der Markierungsdichte mit MedMentions vergleichbar sein. Um einen großen Nutzen des Korpus zu erreichen, soll das UMLS in seiner Gesamtheit möglichst breit abgedeckt sein. Die gewählte Methodik zur Erstellung der Datensätze wird in dieser Arbeit als ein Ensemble-Ansatz bezeichnet. Sie sieht vor, dass zuerst mehrere biomedizinische Entity-Linker zum Extrahieren von UMLS-Entitäten auf die Texte des Datensatzes angewendet und anschließend deren Ausgabewerte zu einem Ergebnis zusammengeführt werden. Die verwendeten Texte sind mehr als 21 Mio. Abstracts aus der PubMed-Baseline von 2021. Bei den Entity-Linkern handelt es sich um scispaCy [26], QuickUMLS [35] und cTAKES [30]. Um die Ausgabewerte zusammenzuführen werden drei denkbare Ensemble-Methoden vorgestellt. Diese nutzen definierte Regeln, um für jede der durch einen der verwendeten Entity-Linker markierten Textstellen zu entscheiden, ob sie im Ensemble ins Ergebnis übernommen wird oder nicht. Die Arbeit soll Aufschluss über die folgenden drei Forschungsfragen geben:

- Welche Kombinationen aus Ensemble-Methode und Entity-Linker-Auswahl führen zu den besten Ergebnissen?
- Sind die Precision-Werte der Evaluation aufgrund von ungleichen Bedingungen bei der Erstellung der Datensätze verfälscht?
- Welchen Mehrwert liefert der vorgestellte Korpus im Vergleich zu anderen Datensätzen?

Um die Forschungsfragen zu beantworten, wird eine umfangreiche Evaluation der Methodik durchgeführt. Zur Beantwortung der ersten Frage werden für alle möglichen Kombinationen aus Ensemble-Methode und Entity-Linkern, verschiedene Kennwerte berechnet und verglichen. Als Goldstandard wird der MedMentions-Datensatz verwendet. Um die zweite Frage zu beantworten, wird anhand eines definierten Regelwerks eine händische Zweitevaluation durchgeführt. Für die dritte Forschungsfrage werden unterschiedliche deskriptive Statistiken erstellt und ausgewertet.

Die vorliegende Arbeit ist in sechs Kapitel aufgeteilt. Kapitel 1 ist die Einleitung. Nach der Einleitung werden in Kapitel 2 verwandte Arbeiten vorgestellt, um das Thema in einen umfassenden Kontext einzubetten. Kapitel 3 erläutert die verwendeten Methoden. Zuerst werden Grundlagen gebildet und die externen Ressourcen erklärt. Danach wird auf die drei Ensemble-Methoden zur Erstellung der Datensätze eingegangen. Kapitel 4 dient der genauen Betrachtung der Experimente. Zunächst werden in einem ersten Evaluations-

schritt anhand des MedMentions-Datensatzes verschiedene Kennwerte berechnet. Anschließend folgt eine händische Zweitevaluation anhand eines vordefinierten Regelwerks. Beim letzten Experiment werden deskriptive Statistiken zu den Datensätzen erstellt und Vergleiche mit MedMentions und PubMedDS gezogen. In Kapitel 5 folgt eine Diskussion der vorliegenden Arbeit. Diskutiert werden die Limitationen des Ansatzes und möglicher zukünftiger Arbeitsbedarf. Außerdem werden Einsatzmöglichkeiten des Korpus aufgezeigt. Das abschließende Kapitel 6 ist eine Zusammenfassung der Arbeit.

Kapitel 2

Verwandte Arbeit

Das folgende Kapitel dient dem Zweck, Literatur von verwandten Themenbereichen zu analysieren und dadurch das Thema dieser Arbeit in einen breiteren Kontext zu stellen. Im ersten Teil werden digitale Ressourcen der Biomedizin vorgestellt. Der daran anschließend Abschnitt handelt von Datensätze mit markierten biomedizinischen Entitäten. Abschließend erfolgt eine Vorstellung mehrerer biomedizinischer Entity-Linker.

2.1 Digitale Ressourcen in der Biomedizin

Die Biomedizin liefert eine große Anzahl an digitalen Ressourcen. Das PubMed ist eine Suchmaschine des National Center for Biotechnology Information (NCBI) für biomedizinische sowie biowissenschaftliche Literatur und beinhaltet Metadaten und Verweise zu mehr als 34 Mio. Veröffentlichungen [24]. Zu den Metadaten einer Veröffentlichung zählen unter anderem Angaben zu den Autoren, der Titel und der Abstract. Die Verweise zeigen auf die Ressourcen von Sammlungen wie dem PubMed Central (PMC) [13]. Das PMC ist ein frei zugängliches Archiv für biomedizinische und biowissenschaftliche Literatur und beinhaltet aktuell ca. 9,1 Mio. Volltextartikel aus mehr als 2.750 wissenschaftlichen Fachzeitschriften.[1] Bookshelf ist ebenfalls eine Sammlung von Volltexten und verfügt über mehr als 10.000 Bücher, Reporte und andere biomedizinische Texte [31].[2] Eine weitere Ressource ist der MEDLINE-Korpus [16]. Dieser beinhaltet keine Volltexte, sondern nur Verweise und Metadaten. MEDLINE besitzt ca. 29 Mio. Einträgen und deckt damit den Großteil des PubMed ab.[3]

[1] Die aktuellen Zahlen zu PMC stammen aus https://www.ncbi.nlm.nih.gov/pmc/. Abgerufen am 12.06.2023

[2] Die aktuellen Zahlen zu Bookshelf wurden durch https://www.ncbi.nlm.nih.gov/books/browse/ ermittelt. Abgerufen am 12.06.2023

[3] Die aktuellen Zahlen zu MEDLINE stammen aus https://www.nlm.nih.gov/medline/medline_overview.html. Abgerufen am 12.06.2023

Das Unified Medical Language System (UMLS) ist ein System, dass es sich zur Aufgabe gemacht hat, die vielen existierenden biomedizinischen Vokabulare zusammenzufassen und zu standardisieren [6]. Beispiele für Vokabulare, die im UMLS vorhanden sind, sind ICD 10, SNOMED CT, NCBI Taxonomy, MeSH und Gene Ontology. Das International Statistical Classification of Diseases and Related Health Problems (ICD) ist ein Vokabular mit medizinischen Diagnosen und wird von der Weltgesundheitsorganisation (WHO) gepflegt [14]. Im ICD werden Erkrankungen über deren Symptomatik und Diagnose definiert. Das ICD 10 ist die zehnte Version. Das SNOMED CT ist eine umfassende multilinguale Terminologie für klinisches Gesundheitswesen [18]. Es stellt einen standardisierten Weg sicher, klinische Daten zu erfassen, und wird verwendet, um elektronische Gesundheitsakten zu erstellen und diese später zu analysieren. Die NCBI Taxonomy ist eine hierarchisch organisierte Namensliste von Organismen aus allen wichtigen Forschungsbereiche die das Leben betreffen [32]. Die Medical Subject Headings (MeSH) sind ein Vokabular, das zur Indexierung biomedizinischer Dokumente verwendet wird [23]. Die Gene Ontology (GO) stellt standardisiertes Wissen über die Funktionen von Genen und Genprodukten zur Verfügung [9]. Die GO definiert dabei Klassen von Genen (die GO-Terme) und verknüpft diese über Relationen miteinander.

2.1.1 Datensätze

Weitere digitale Ressourcen sind Datensätze. MedMentions ist ein händisch annotierter Datensatz mit Goldstandard Qualität [25]. Der Datensatz beinhaltet 4.392 Texte, in denen Textstellen von Experten markiert und mit UMLS-Entitäten verlinkt sind. Der MedMentions-Datensatz spezialisiert sich nicht auf einen Teil der Biomedizin, sondern besitzt eine breite Streuung an Konzepttypen. Er beinhaltet beispielsweise Konzepte der Gruppen Krankheiten, Behandlungsformen oder Chemikalien. Andere Datensätze spezialisieren sich oft auf eine spezielle Art von Konzepten. Der NCBI-Korpus ist ein spezialisierter Datensatz für Entitäten aus der Kategorie Krankheiten [12]. Er wurde durch eine Gruppe von zwölf Experten annotiert und beinhaltet 2.783 Sätze. Der BC5CDR spezialisiert sich auf den Zusammenhang zwischen Chemikalien und Krankheiten [20]. Der Datensatz ist ebenfalls händisch annotiert und besitzt 1.500 Texte, in denen MeSH-Markierungen gesetzt wurden. Der BC4GO ist ein Korpus für Begriffe aus der GO [4]. Acht GO-Experten haben 5.000 Textpassagen untersucht und darin insgesamt 1.356 GO-Konzepte markiert. Der ShARe-Korpus wurde für das ShARe/CLEF eHealth 2013 Evaluation Lab entworfen und enthält anonymisierte klinische Notizzettel [29]. In diesen wurden von trainierten Experten Konzepte des SNOMED CT markiert. Der PubMedDS-Datensatz ist im Zuge einer Forschungsarbeit zur Verbesse-

rung des allgemeinen medizinischen Entity-Linking durch die Vorhersage von semantischen Typen und der Nutzung sehr großer Datensätze entstanden [36]. Der Datensatz ist im Vergleich zu anderen bekannten Datensätzen sehr groß und breit gestreut, mit über 13 Mio. Texten und mehr als 56 Mio. markierten UMLS-Konzepten. Zur Erstellung der Markierungen wurde ein Ansatz mit Distant Supervision gewählt. Dieser sieht vor, ein biomedizinisches NER-System auf PubMed-Abstracts durchzuführen und die Korrektheit der Markierungen mittels der zu den Texten vorhandenen MeSH-Terme zu überprüfen. Dadurch ist es möglich, automatisiert UMLS-Entitäten zu markieren und einen großen Datensatz zu erstellen.

2.2 Entity-Linking

Die vorgestellten digitalen Ressourcen sind eine ideale Grundlage für die Entwicklung neuer datengetriebener Anwendungen und Systeme. Ein Bereich dieser Anwendungen ist der des biomedizinischen Entity-Linking. Dabei handelt es sich um Programme die in Texten biomedizinische Entitäten finden und diese mit Entitäten eines biomedizinischen Vokabulars (z.B. UMLS, MeSH, GO) verlinken [39]. scispaCy ist ein Erweiterungspaket für das NLP Toolkit spaCy und verfolgt einen K Nearest Neighbours-Ansatz, um für eine identifizierte Textstelle die K-wahrscheinlichsten Entitäten aus einer Wissensdatenbank zu ermitteln [26]. Mögliche Wissensdatenbanken bei scispaCy sind UMLS, MeSH, RxNorm, GO und Human Phenotype Ontology. cTAKES ist ein Pipeline-basiertes, modulares System von Apache zur Textanalyse und Wissensextraktion [30]. cTAKES kombiniert regelbasierte Aspekte mit maschinellem Lernen, um UMLS-Entitäten aus Texten zu extrahieren. QuickUMLS ist ein Entity-Linker für das Extrahieren von UMLS-Entitäten mit einem besonderen Augenmerk auf Geschwindigkeit [35]. QuickUMLS nutzt dafür den Algorithmus CPMerge, der ein unscharfes Wörterbuch-Matching ermöglicht [27]. MetaMap ist ein Entity-Linker von der National Library of Medicine (NLM) und Dr. Alan Aaronson [3]. Der Ansatz von MetaMap basiert auf symbolic NLP und anderen linguistischen Techniken. MetaMap-Lite ist eine Neuimplementierung der grundlegenden Funktionen von MetaMap mit einem Fokus auf Datenverarbeitung in Echtzeit und Performance [11]. MedLinker nutzt zum Identifizieren der Textstellen mit UMLS-Konzepten ein NER-System mit einer BiLSTM CRF-Architektur [22]. Die Zuordnung der Textstellen zu den Konzepten erfolgt dann mit einem kombinierten Ansatz aus Wörterbuch-Matching und kontextbasiertem Matching. BioBART ist die Adaption von Bidirectional and Auto Regressive Transformers (BART) [19] auf die biomedizinische Domäne [37]. Die Spezialisierung wird erreicht, in-

dem das Modell auf PubMed Abstracts vortrainiert wird. ArboEL erstellt eine gewurzelte Baumstruktur (engl. Arborescence), um im Trainingsschritt die Beziehungen der Markierungen untereinander zu lernen und dadurch das Linking zu verbessern [1]. Dadurch erreicht ArboEL die aktuell höchste Genauigkeit beim Entity-Linking auf den MedMentions-Datensatz in PapersWithCode.[4]

[4] Aktuelles Ranking des Entity-Linkings auf dem MedMentions-Datensatz von PapersWithCode ist unter https://paperswithcode.com/sota/entity-linking-on-medmentions zu finden. Abgerufen am 12.06.2023

Kapitel 3

Methoden

In diesem Kapitel werden die Methoden zur Erstellung des Korpus erläutert. Den Einstieg macht die Grundlagenbildung. Dies ist nötig, damit die weiteren Abschnitte dieser Arbeit ohne spezifisches Vorwissen verstanden werden können. Zu den Grundlagen zählen das UMLS, das PubMed, der MedMentions-Datensatz, der PubMedDS-Datensatz und Entity-Linking für UMLS-Konzepte. Nach der Grundlagenbildung folgt die Erklärung der Methodik die verwendet wird um den Korpus zu erstellen. Dabei wird zuerst auf die allgemeine Methode eingegangen. Danach werden Ensemble-Methoden vorgestellt, die es ermöglichen die separaten Ergebnisse verschiedener Entity-Linker in einem Ensemble zusammenzuführen. Abschließend werden Details zur technischen Umsetzung erläutert.

3.1 Grundlagen

Zu den Grundlagen zählen das UMLS, das PubMed, der MedMentions-Datensatz, der PubMedDS-Datensatz und Entity-Linking für UMLS-Konzepte. Die Klärung der Grundlagen ist wichtig für die Verständlichkeit der später folgenden Teile dieser Arbeit.

3.1.1 UMLS

Das Unified Medical Language System (UMLS) ist eine Sammelstelle von biomedizinischem Vokabular und wurde erstmals 1989 durch die US National Library of Medicine (NLM) veröffentlicht [6]. Es gleicht unterschiedliche Terminologien verschiedenster Ressourcen aneinander an und vereinheitlicht diese. Dadurch soll die Arbeit mit unterschiedlichen biomedizinischen Datenquellen erleichtert werden. Um Zugriff auf die Ressourcen des UMLS zu erhalten, wird

eine Lizenz benötigt.[1] Die Lizenz wird nur an Einzelpersonen vergeben und ist kostenfrei. Das UMLS löst zwei Probleme, die einer effektiven Informationsbeschaffung für biomedizinische Anwendungen häufig im Wege stehen [21]. Das erste ist die Menge an unterschiedlichen Begriffen, die in verschiedenen Vokabularen verwendet werden, um dasselbe Konzept zu benennen. Im UMLS werden die unterschiedlichen Bezeichnungen gesammelt und in einer Entität mit einem einzigartigen Identifikator zusammengefügt. Das zweite Problem ist die Verteilung von biomedizinischen Informationen in viele verschiedenen Datenbanken und Systemen. Die Informationen werden durch das UMLS gesammelt und in einem System gebündelt. Dadurch wird die Informationsbeschaffung in der Biomedizin erleichtert. In einer Umfrage des Jahres 2018 wurden 5.043 Nutzer befragt, welchen Zweck das UMLS für sie erfüllt [2]. Die meisten Personen (51%) haben angegeben, das UMLS zu verwenden um Konzepte, Relationen und anderes Wissen aus Texten zu extrahieren. Die zweithäufigste Verwendung (49%) findet das UMLS, beim Zusammenführen mehrerer Terminologien. Des Weiteren benutzen 29% der Befragten das UMLS um spezifische Terminologien (z.B. MedRRA, MeSH, NDF-RT) zu extrahieren, 19% verwenden das UMLS im Rahmen eines Information Retrieval Systems und ebenfalls 19% zur Entwicklung eigener lokaler Terminologien. Diese Umfrageergebnisse zeigen den Mehrwert, den das UMLS für die Biomedizin hat.

Komponenten des UMLS

Es gibt drei Hauptkomponenten des UMLS [6]. Die erste ist das Metathesaurus. Es beinhaltet miteinander verknüpfte biomedizinische Konzepte. Die zweite Komponente ist das Semantic Network und beschreibt semantische Typen und deren Beziehungen zueinander. Zuletzt gibt es das SPECIALIST Lexicon und Lexical Tools. Diese beiden Teilsysteme dienen als lexikalische Grundlage für viele verschiedene NLP-Prozesse. Die letzte Komponente ist für die vorliegende Arbeit nicht relevant und wird daher nicht weiter behandelt. Das Metathesaurus und das Semantic Network werden nachfolgend genauer beschrieben. Um die Bestandteile des UMLS besser verstehen zu können, folgt auf die Beschreibung der Komponenten ein ausführliches Beispiel.

Das **Metathesaurus** ist eine große, multilinguale Datenbank mit Informationen über biomedizinische Entitäten. Entitäten werden im Metathesaurus auch als Konzepte bezeichnet. Die enthaltenen Informationen sind beispielsweise Listen synonymer Namen für ein Konzept, die Beziehungen der Konzepte zueinander oder Begriffsdefinitionen. Die Datensammlung des Metathesaurus

[1] Alle Informationen zur UMLS-Lizenz sind unter https://www.nlm.nih.gov/research/umls/index.html verfügbar. Abgerufen am 12.06.2023

entsteht durch die Zusammenführung von elektronischen Versionen vieler biomedizinischer Quellvokabulare. Bei den Quellen handelt es sich um Sammelwerke, Klassifikationen, Begriffslisten, Lexika und viele weitere Ressourcen. Es stehen mehr als 100 Quellvokabulare zur Verfügung und je nach Anwendungsfall kann jedes davon in das Metathesaurus mit aufgenommen bzw. abgewählt werden. Dadurch wird eine effektive und personalisierte Nutzung ermöglicht. Das Metathesaurus ist als Wissensgraph organisiert. Dieser beinhaltet biomedizinische Konzepte als Knoten und deren Beziehungen untereinander als Kanten. Ein UMLS-Konzept wird durch die Zusammenführung von synonymen Begriffen, Namen und Definitionen aus den verschieden Quellvokabularen gebildet. Jedes Konzept hat einen einzigartigen Identifikator (die CUI), einen primären Namen und eine Liste von Pseudonymen. Beispielsweise besitzt das Konzept mit der CUI C0030827 den primären Namen 'penicillin G' und eine Liste mit den Pseudonymen 'benzylpenicillin', 'PG' und 'Free Penicillin II'. Die Beziehungen der Konzepte untereinander werden meist aus den Quellvokabularen geerbt oder durch die UMLS-Entwickler nachträglich hinzugefügt und sind entweder hierarchisch (z.B. 'isa', 'part of', 'broader concept') oder assoziativ (z.B. 'location of', 'caused by'). Als Beispiel für eine Beziehung ist das Konzept 'penicillin G' in Form einer 'broader concept'-Relation mit dem Konzept 'antibiotics' verbunden. Das Konzept 'antibiotics' ist dem Konzept 'penicillin G' also hierarchisch übergeordnet. Durch das Zusammenspiel aus Konzepten und Beziehungen entsteht bei Auswahl aller Quellvokabulare ein großes Graphenmodell mit aktuell ca. 3,3 Mio. Konzepten.[2]

Ein weiterer Bestandteil des UMLS sind semantische Typen. Die semantischen Typen sind allgemeine Kategorien, welche das UMLS in semantisch zueinander passende Gruppen gliedern, und sind in einem Netzwerk – dem **Semantic Network** - mit hierarchischer Struktur organisiert. Genau wie die UMLS-Konzepte, haben auch die semantischen Typen einen einzigartigen Identifikator (die TUI). Jedem UMLS-Konzept ist mindestens ein semantischer Typ zugeordnet. Zum Beispiel sind dem Konzept 'penicillin G' die beiden semantischen Typen 'Antibiotic' mit der TUI T195 und 'Organic Chemical' mit der TUI T109 zugeordnet. Die hierarchische Struktur der semantischen Typen beginnt mit den beiden allgemeinsten Typen 'Entity' und 'Event', welche jeweils den Beginn eines eigenen Teilbaums darstellen. Diese Teilbäume werden mittels der 'isa'-Beziehung immer weiter verfeinert je tiefer die Ebene in den Bäumen verfolgt wird. Der semantische Typ 'Organic Chemical' mit der TUI T109 ist beispielsweise auf der sechsten Hierarchieebene des Teilbaums 'Entity' zu finden und ist mittels der 'isa'-Relation mit dem semantischen Typen 'Chemical

[2]Die aktuellen Zahlen zum Metathesaurus sind unter https://uts.nlm.nih.gov/uts/umls/home zu finden. Abgerufen am 12.06.2023

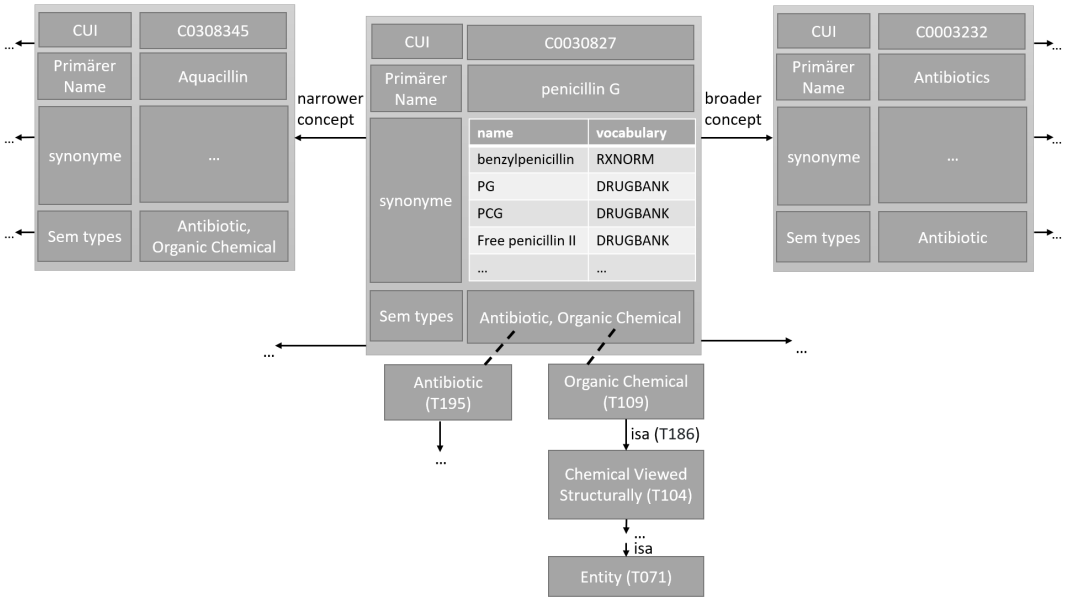

Abbildung 3.1: An diesem Beispiel ist zu sehen, wie die Konzepte des Metathesaurus und Semantic Network zusammenspielen. Dargestellt ist das Konzept C0030827. Das Konzept besitzt eine CUI, einen primären Namen, eine Liste von Synonymen aus verschiedenen Vokabularen und eine Liste von semantischen Typen. Das Konzept C0030827 besitzt Beziehungen zu anderen Konzepten im Metathesaurus. Außerdem sind die semantischen Typen Teil des Semantic Network und stehen in hierarchischer Beziehung zu anderen semantischen Typen.

Viewed Structurally' mit der TUI T104 verbunden, der sich in der Hierarchie auf dem fünften Level befindet. Neben der 'isa'-Beziehung gibt es noch viele weitere Beziehungen, die unter dem Teilbaum 'associated_with' zusammengefasst und ebenfalls hierarchisch geordnet sind. Die in dieser Arbeit verwendete UMLS-Version 2022AA gliedert das Semantic Network in insgesamt 127 semantische Typen. Davon gehören 92 zum Typen 'Entity' und 35 zum Typen 'Event'. Außerdem gibt es 54 Relationen, wovon eine die 'isa'-Relation ist und 53 dem Typen 'associated_with' untergeordnet sind.[3]

Abbildung 3.1 zeigt ein **ausführliches Beispiel** zu den oben beschriebenen Komponenten des UMLS. Dies dient als Hilfe, um ein grundlegendes Verständnis für das UMLS und dessen Prinzipien zu schaffen. Dadurch soll Verwirrungen bei der Benutzung des UMLS im weiteren Verlauf dieser Arbeit verhindert werden. Das Konzept mit der CUI C0030827 besitzt den primären

[3]Die angegeben Zahlen zum Semantic Network sind durch https://uts.nlm.nih.gov/uts/umls/semantic-network/root ermittelt worden. Abgerufen am 12.06.2023

Namen ‘penicillin G’ und eine Liste von Pseudonymen (‘benzylpenicillin’, ‘PG’, ‘PCG’, usw.). Außerdem sind dem Konzept die semantischen Typen ‘Antibiotic’ mit der TUI T195 und ‘Organic Chemical’ mit der TUI T109 zugeordnet. Es muss immer mindestens ein semantischer Typ zugeordnet sein. Das Konzept C0030827 steht in Beziehung zum Konzept C0003232 mit dem primären Namen ‘Antibiotics’. Bei der Beziehung handelt es sich um eine ‘broader concept’-Relation. ‘penicillin G’ ist also eine spezielle Form eines ‘Antibiotics’. Gleichzeitig besteht eine ‘narrower concept’-Beziehung zum Konzept C0308345 mit dem primären Namen ‘Aquacillin’. ‘penicillin G’ ist also zugleich auch eine Übergruppierung und ‘Aquacillin’ kann dieser Gruppe zugeordnet werden. Zusätzlich gibt es noch viele weitere Beziehungen zu verwandten Konzepten, die in diesem Beispiel durch weitere von den Konzepten wegführende Pfeile dargestellt werden sollen. Die semantischen Typen des Konzepts sind Teil des Semantic Networks und besitzen eigene Beziehungen. Der semantische Typ ‘Organic Chemical’ besitzt zum Beispiel eine ‘isa’-Beziehung zum semantischen Typen ‘Chemical Viewed Structurally’. ‘Organic Chemical’ ist also eine Untergruppierung der Gruppe ‘Chemical Viewed Structurally’. Diese ‘isa’-Beziehung kann hierarchisch weiterverfolgt werden, bis mit ‘Entity’ die oberste Ebene erreicht ist. Das Beispiel verdeutlicht, dass das Konzept C0030827 Teil eines großen Netzwerks mit vielen Konzepten und Beziehungen, sowohl zwischen den Konzepten als auch zwischen den semantischen Typen, ist.

3.1.2 PubMed

Das PubMed ist eine frei nutzbare Suchmaschine des National Center for Biotechnology Information (NCBI) und liefert Verweise und Links zu biomedizinischer und andere biowissenschaftlicher Literatur [24]. Es beinhaltet aktuell mehr als 34 Mio. Abstract-Texte und Zitationen.[4] Das PubMed liefert keine Volltexte, sondern lediglich Metadaten. Dazu zählen unter anderem die Autoren, der Abstract, der Titel und Verlinkungen zu Volltexten. Zudem beinhaltet ein Suchergebnis in PubMed meistens Medical Subject Headings (MeSH) [23]. Sie dienen als Stichwortverzeichnis und zur Indexierung der Dokumente. Durch die Benutzung von MeSH-Termen wird ein konsistenter Weg sichergestellt, um bei einer Suchanfrage alle relevanten Artikel als Suchergebnisse zu liefern, selbst wenn die Artikel unterschiedliche Begrifflichkeiten nutzen.

Das PubMed liefert Zitationen zu Ressourcen von drei großen Datenquellen. Diese sind MEDLINE, PubMed Central (PMC) und Bookshelf. Die größte Quelle ist der MEDLINE-Korpus. Er besteht hauptsächlich aus Zitationen und

[4] Die aktuelle Zahlen zum PubMed sind aus https://pubmed.ncbi.nlm.nih.gov/about/ entnommen. Abgerufen am: 12.06.2023

beinhaltet somit keine Volltexte. MEDLINE ist die einzige der drei Quellen, die MeSH-Terme verwendet und liefert Metadaten wie beispielsweise den Titel und Abstract. Die zweitgrößte Ressource ist PubMed Central (PMC). PMC ist ein Archiv für aktuelle und historische biomedizinische Volltexte. Die kleinste Ressource ist Bookshelf. Dabei handelt es sich ebenfalls um ein Volltextarchiv. Dieses beinhaltet aber vor allem ganze Bücher, Reporte, Datenbanken und andere Dokumente mit biomedizinischem Hintergrund. Die Inhalte der drei Datenquellen sind alle frei zugänglich. Für die vorliegende Arbeit ist vor allem der MEDLINE-Korpus relevant, da er die Datengrundlage für den praktischen Teil liefert.

3.1.3 MedMentions

Der MedMentions-Datensatz ist ein manuell annotierter Datensatz und wurde für Modelle zur Erkennung biomedizinischer Konzepte erstellt [25]. Vergleichbare Datensätze wie der NCBI [12] oder der BC4GO [4] sind in der Regel auf bestimmte Konzepttypen limitiert - z.B. Krankheiten bei NCBI; Gene bei BC4GO - und besitzen maximal ein paar Tausend mit Entitäten annotierte Textstellen. Deshalb haben die Autoren bei der Erstellung des MedMentions-Korpus auf eine breite Streuung unterschiedlicher biomedizinischer Fachbereiche und Konzepttypen sowie auf eine signifikante Vergrößerung der bisherigen Korpusgrößen geachtet. Enthalten sind 4.392 Abstracts von zufällig ausgewählten PubMed-Artikeln, welche im Jahr 2016 erschienen sind. In diesen Texten wurden von einem Expertenteam umfänglich UMLS-Konzepte markiert. Für jede identifizierte Textstelle mussten die Experten die 2017AA (full) Version des UMLS Metathesaurus durchsuchen und sie mit dem ihrer Meinung nach am besten passendem UMLS-Konzept verknüpfen. Dabei war vorgegeben, immer das spezifischste UMLS-Konzept zu wählen und keine sich überlappende Markierungen zu setzen.

Nach dem Annotieren durch die Experten wurden die Konzeptmarkierungen weiter eingeschränkt und es wurden einige Konzepte aus dem Datensatz herausgefiltert. Als erstes wurden alle annotierten Konzepte entfernt, deren zugeordnete semantische Typen ausschließlich auf Level 1 oder 2 der Hierarchie stehen, da diese als zu allgemein betrachtet wurden. Aus den übrig gebliebenen semantischen Typen wurden dann die 21 relevantesten ausgewählt. Dabei wurde sowohl auf die Relevanz in der Biomedizin als auch auf die Häufigkeit in den annotierten Textstellen geachtet. Eine genau Auflistung der 21 relevanten semantischen Typen ist in Tabelle 3.1 zu finden. Alle Konzepte, die nicht

[5]Tabelle übernommen aus Mohan und Li [25].

Tabelle 3.1: Eine Aufstellung der 21 relevanten semantischen Typen, die bei der Erstellung von MedMentions erlaubt waren. Konzepte, die nicht mindestens einem der markierten semantischen Typen oder einem untergeordnetem Element zugeordnet sind, kommen in MedMentions nicht vor.[5]

Typ Name	TUI	hier. Ebene
Event	T051	1
Activity	T052	2
Behavior	T053	3
Social Behavior	T054	4
Individual Behavior	T055	4
Daily or Recreational Activity	T056	3
Occupational Activity	T057	3
Health Care Activity	**T058**	**4**
Research Activity	**T062**	**4**
Governmental or Regulatory Activity	T064	4
Educational Activity	T065	4
Machine Activity	T066	3
Phenomenon or Process	T067	2
Injury or Poisoning	**T037**	**3**
Human-caused Phenomenon or Process	T068	3
Environmental Effect of Humans	T069	4
Natural Phenomenon or Process	T070	3
Biologic Function	**T038**	**4**
Entity	T071	1
Physical Object	T072	2
Organism	T001	3
Virus	**T005**	**4**
Bacterium	**T007**	**4**
Archaeon	T194	4
Eukaryote	**T204**	**4**
Anatomical Structure	**T017**	**3**
Manufactured Object	T073	3
Medical Device	**T074**	**4**
Research Device	T075	4
Clinical Drug	T200	4
Substance	T167	3
Body Substance	**T031**	**4**
Chemical	**T103**	**4**
Food	**T168**	**4**
Conceptual Entity	T077	2
Organism Attribute	T032	3
Clinical Attribute	**T201**	**4**
Finding	**T033**	**3**
Idea or Concept	T078	3
Temporal Concept	T079	4
Qualitative Concept	T080	4
Quantitative Concept	T081	4
Spatial Concept	**T082**	**4**
Functional Concept	T169	4
Body System	**T022**	**5**
Occupation or Discipline	T090	3
Biomedical Occupation or Discipline	**T091**	**4**
Organization	**T092**	**3**
Group	T096	3
Professional or Occupational Group	**T097**	**4**
Population Group	**T098**	**4**
Family Group	T099	4
Age Group	T100	4
Patient or Disabled Group	T101	4
Group Attribute	T102	3
Intellectual Product	**T170**	**3**
Language	T171	3

mindestens einem der 21 dort markierten semantischen Typen oder einem untergeordnetem Element dieser zugeordnet sind, wurden entfernt. Zum Schluss wurden 18 Quellvokabulare - basierend auf deren Relevanz in der Biomedizin - festgelegt. Konzepte, die nicht in mindestens einem der 18 Vokabulare vorkommen, wurden entfernt. Nach all diesen Schritten besitzt MedMentions 4.392 einzigartige UMLS-Konzepte die in insgesamt 352.496 Textstellen markiert sind. Es ist wichtig, sich bei der Nutzung des Datensatzes stets der getroffenen Einschränkungen bewusst zu sein. Je nach Anwendung sollten dann gegebenenfalls Schritte durchgeführt werden, um diese Einschränkungen zu berücksichtigen.

3.1.4 PubMedDS

Der PubMedDS-Datensatz entstand im Rahmen einer Forschungsarbeit von Vashishth et al. [36] und umfasst ca. 13 Mio. Dokumente aus dem PubMed. Das Ziel der Forschungsarbeit bestand darin, das allgemeine Entity-Linking in der biomedizinischen Domäne zu verbessern. Das soll durch das Verwenden eines Moduls zur Vorhersage der semantischen Typen und der Einführung von sehr großen Trainingsdatensätzen ermöglicht werden. Einer der darin vorgestellten Trainingsdatensätze ist PubMedDS. Zur Erstellung nutzten die Autoren Distant Supervision, um anhand der für die PubMed-Artikel vorhandenen MeSH-Terme automatisiert UMLS-Konzepte zu markieren.

Abbildung 3.2 zeigt das Vorgehen grafisch auf. Zunächst wird ein NER-System zur Extraktion biomedizinischer Entitäten auf die Abstracts von PubMed-Artikeln angewendet. Die dadurch erhaltenen Konzepte werden durch die in PubMed angegebenen MeSH-Terme gegengeprüft und es werden nur die Konzepte in den Datensatz übernommen, die sowohl durch das NER als auch in den MeSH-Termen gefunden werden. Das gewählte Verfahren ermöglicht es, auf einfache Art und Weise einen sehr großen Datensatz mit vielen markierten UMLS-Konzepten zu erstellen. Allerdings bringt es auch Einschränkungen für den Datensatz mit sich. Erstens führt der beschriebene Weg dazu, dass nur UMLS-Konzepte vorhanden sind, welche auch im MeSH-Vokabular vorkommen. Das entspricht mit ca. 450.000 Konzepten in etwa 10% des gesamten UMLS 2022AA.[6] Außerdem werden, durch die Überprüfung mit den MeSH-Termen ein großer Teil der durch das NER erkannten Konzepte wieder verworfen. Beide Punkte führen dazu, dass im Vergleich mit MedMentions bei

[6]Die Zahl der 450.000 Konzepte wurde bei der Erstellung eines UMLS-subset ermittelt. Eine Anleitung zur Erstellung des UMLS-subset ist unter https://www.nlm.nih.gov/research/umls/implementation_resources/metamorphosys/help.html zu finden. Abgerufen am 13.06.2023

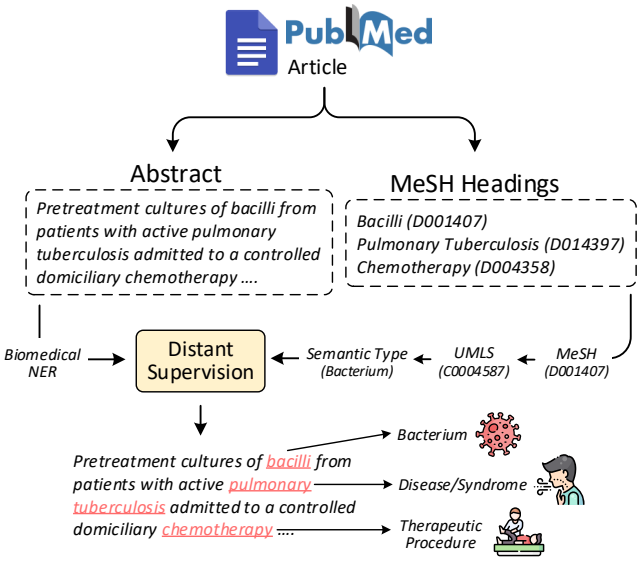

Abbildung 3.2: Methode zur Erstellung von PubMedDS. Für jeden PubMed Artikel wird auf den Abstract biomedizinisches NER durchgeführt, um alle möglichen Entitäten zu identifizieren. Anschließend werden mittels Distant Supervision und den MeSH-Termen des Artikels die UMLS-Entitäten extrahiert, die in den Datensatz übernommen werden.[7]

einer Precision von 90,3%, ein Recall von gerade einmal 5,3% erreicht wird. Daher ist der Datensatz vor allem für Anwendungen geeignet, die eine hohe Precision erfordern und für die der Recall zweitrangig ist. PubMedDS dient aufgrund seiner Größe und automatisierten Methodik als Inspiration für das in dieser Arbeit vorgestellten Datensatzkonzeptes.

3.1.5 Entity-Linking für UMLS-Konzepte

Entity-Linking bezeichnet einen Prozess, bei dem in einem unstrukturierten Dokument erwähnte Entitäten mit Entitäten aus einer Wissensdatenbank verknüpft werden [39]. Im vorliegenden Fall ist die Wissensdatenbank das UMLS und die gesuchten Entitäten sind UMLS-Konzepte. Modelle, die Entity-Linking durchführen, werden in dieser Arbeit Entity-Linker genannt. Im Folgenden werden die Entity-Linker scispaCy, cTAKES und QuickUMLS vorgestellt.

[7] Grafik übernommen von Vashishth et al. [36].

Tabelle 3.2: Auflistung und Vergleich aller scispaCy Pipeline-Modelle.[9]

Modell	Vokabular Größe	Zusatz
en_core_sci_sm	ca. 100.000 Einträgen	-
en_core_sci_md	ca. 360.000 Einträgen	50.000 Wortvektoren
en_core_sci_lg	ca. 785.000 Einträgen	600.000 Wortvektoren
en_core_sci_scibert	ca. 785.000 Einträgen	SciBERT-Transformer

scispaCy

scispaCy[8] ist ein Erweiterungspaket für das NLP-Toolkit spaCy [26]. spaCy ist eine Python-basierte Bibliothek und stellt eine große Anzahl an unterschiedlichen Tools zur Textverarbeitung zur Verfügung. Aufgrund der Beliebtheit von spaCy und dessen Modellen und Methoden haben sich die scispaCy-Entwickler dafür entschieden, darauf aufbauend eine NLP-Bibliothek für die biomedizinische Domäne zu erstellen. Für die Erstellung von scispaCy wurden spaCy-Modelle für Part of speech (POS) Tagging, Dependency Parsing und Named Entity Recognition (NER) mit biomedizinisch relevanten Datensätzen - darunter MedMentions - neu trainiert. Außerdem wurde das Tokenisierungsmodul von spaCy mit speziellen Regeln erweitert. Für das Entity-Linking wird ein K Nearest Neighbours-Verfahren verwendet. Dadurch werden für eine Textstelle die K-wahrscheinlichsten Entitäten aus der ausgewählten Wissensdatenbank zurückgegeben. Mögliche Wissensdatenbanken sind UMLS, MeSH, RxNorm, GO und Human Phenotype Ontology. Ein weiterer Bestandteil des Entity-Linking mit scispaCy ist eine Komponente zur Auflösung von Abkürzungen. Ist diese Option aktiviert, werden entdeckte Abkürzungen mittels des Algorithmus von Schwartz und Hearst [33] vor der Generierung der Linking-Kandidaten durch deren Langform ersetzt.

scispaCy stellt acht Modelle zur Verfügung. Vier davon sind vollwertige spaCy-Pipelines für biomedizinische Textverarbeitung und die anderen vier sind eigenständige, auf unterschiedlichen Korpora trainierte, NER-Modelle. Die vier Pipelines unterscheiden sich durch die Größe des Vokabulars und die Verwendung von Wortvektoren. Mit 'en_core_sci_scibert' gibt es außerdem eine Pipeline, die ein SciBERT-basiertes [5] Transformer-Modell nutzt. Tabelle 3.2 stellt die vier für diese Arbeit interessanten Pipelines gegenüber. Das Entity-Linking ist jeweils als Komponente der Pipeline-Modelle verwendbar.

[8] Eine Dokumentation zu scispaCy ist unter https://github.com/allenai/scispacy zu finden. Abgerufen am 12.06.2023

[9] Informationen der Tabelle stammen aus https://github.com/allenai/scispacy#available-models . Abgerufen am 12.06.2023

cTAKES

Das clinical Text Analysis and Knowledge Extraction System (cTAKES)[10] ist ein modulares System mit unterschiedlichen Pipeline-Komponenten [30]. Es ist unter einer open source Apache-Lizenz verfügbar und wurde unter Verwendung von der Apache Unstructured Information Management Architecture (UIMA) und Java 1.5 implementiert. Die Pipeline-Komponenten können individuell und den eigenen Anforderungen entsprechend zusammengefügt werden. Für ein Eingabedokument wird die Pipeline dann Schritt für Schritt durchlaufen und so der gewünschte Output generiert. cTAKES kombiniert regelbasierte Aspekte mit maschinellem Lernen. Die Default Clinical Pipeline ist die Standartpipeline zur Extraktion medizinischer Entitäten. In einem ersten Schritt werden die Satzgrenzen ermittelt. Dafür wird die Satzgrenzenerkennung von OpenNLP[11] erweitert, um für Punkte, Fragezeichen oder Ausrufezeichen vorherzusagen, ob es sich um das Ende eines Satzes handelt. Anschließend wird für jeden Satz eine Tokenisierung durchgeführt. Der Tokenizer besteht aus zwei Subkomponenten. Die erste teilt den Text eines Satzes an Leerzeichen und Interpunktionen auf. Die zweite Subkomponenten ist eine kontextabhängige Tokenisierung und fügt falsch aufgeteilte Token wieder zusammen. Beispielsweise wird das Datum 29.03 aufgrund des '.' zunächst in zwei Token aufgeteilt. Durch die kontextabhängige Tokenisierung werden die beiden Token wieder zusammengefügt. Nach der Tokenisierung wird ein POS Tagging durchgeführt, um die Wortarten und Informationen über die Satzzusammenstellung zu erhalten. Danach wird der Shallow Parser verwendet, um Chunking anzuwenden und zusammengehörende Teile eines Satzes zu gruppieren. Im letzten Schritt wird dann Entity-Linking durchgeführt. Dabei werden durch einen Wörterbuch-Lookup-Algorithmus in den durch das Chunking festgelegten Nominalgruppen biomedizinische Entitäten identifiziert und mit UMLS-Konzepten verknüpft.

QuickUMLS

QuickUMLS[12] ist ein schneller, nicht überwachter Algorithmus für das unscharfe Wörterbuch-Matching, der zur Extraktion medizinischer Entitäten in unstrukturierten Texten verwendet wird [35]. Es handelt sich dabei um eine frei nutzbare Python-Implementierung. QuickUMLS ist im Vergleich zu anderen Entity-Linkern auf Geschwindigkeit optimiert und eignet sich daher gut für

[10] Eine Dokumentation zu cTAKES ist unter https://github.com/apache/ctakes/wiki zu finden. Abgerufen am 12.06.2023

[11] OpenNLP ist ein NLP Bibliothek von Apache. Informationen sind unter https://opennlp.apache.org/ zu finden. Abgerufen am 12.06.2023

[12] Eine Dokumentation zu QuickUMLS ist unter https://github.com/Georgetown-IR-Lab/QuickUMLS zu finden. Abgerufen am 12.06.2023

die Arbeit mit großen Datensätzen. Das System ist in der Lage, ein Dokument mit ca. 1.000 Token in 500 bis 1.000 ms zu bearbeiten und ist dadurch bis zu 135-mal schneller als cTAKES.

QuickUMLS nutzt den CPMerge-Algorithmus [27], um die Komplexität des Matching-Schrittes zu verringern und dadurch Rechenzeit zu sparen. Das erreicht CPMerge, indem das Wörterbuch durch einen Inverted Index repräsentiert wird. Der Inverted Index speichert Features. Die Features sind Trigramme aller Wörterbucheinträge. Jedem Feature werden alle Strings des Wörterbuchs zugeordnet, die dieses Feature besitzen. Wird für einen neuen Eingabestring nun nach möglichen Treffern im Wörterbuch gesucht, wird zunächst ermittelt, wie viele Features minimal und maximal einem Eintrag im Wörterbuch zugeordnet sein müssen, um eine vorgegebene Grenze bei einer vorher gewählten Ähnlichkeitsfunktion zu erreichen. Die Ähnlichkeitsfunktion bei QuickUMLS ist die Jaccard Sim. In die Ergebnisliste der möglichen Wörterbuch-Matches gehen nur die Einträge ein, deren Anzahl an übereinstimmenden Features zwischen den errechneten Minimal- und Maximal-Werten liegt.

Der QuickUMLS-Algorithmus beginnt mit einer Tokenisierung des gesamten Dokuments mittels spaCy. Die erhaltenen Token werden nun nach und nach durchlaufen. Zunächst werden für jedes Token mittels eines Sliding Windows mehrere Sequenzen von Token festgelegt, die auf mögliche UMLS-Entitäten untersucht werden sollen. Für ein Sliding Window der Größe drei werden beispielsweise für jedes Token drei Sequenzen erstellt. Die erste Sequenz beinhaltet nur das Token selbst. Die zweite Sequenz beinhaltet das Token selbst und das Token, das danach kommt. Die dritte Sequenz beinhaltet schließlich das Token selbst und die zwei Token, die danach kommen. Anhand definierter Heuristiken wird dann für jede Tokensequenz überprüft, ob sie zulässig ist. Zum Beispiel darf sich die Tokensequenz nicht über mehrere Sätze erstrecken oder das erste Token in der Sequenz darf keine Interpunktion sein. Anschließend wird CPMerge genutzt, um aus einem UMLS-Wörterbuch für jede Tokensequenz die Strings zu finden, welche bei Ermittlung der Jaccard Sim größer oder gleich einem Grenzwert α sind. Die Treffer werden dann als Liste zurückgegeben.

3.2 Erstellung des Korpus

Nachdem die Grundlagen eingeführt wurden, folgt im nächsten Abschnitt die Erklärung, wie der neue Korpus erstellt wird. Zunächst wird die Methode erläutert. Anschließend wird auf die Bildung der Ensemble-Methoden eingegangen. Zum Ende wird gezeigt, wie die Theorie technisch umgesetzt werden kann.

3.2.1 Methode

Bei der Erstellung des PubMedDS wurde zu Gunsten einer hohen Precision ein niedriger Recall in Kauf genommen. Um den Datensatz für beispielsweise die Erstellung einer Suchmaschine nutzen zu können, sollte aber der Recall um einiges erhöht werden. Deshalb wird die Methodik zur Erstellung von PubMedDS herangezogen und an dieses Bedürfnis angepasst.

Abbildung 3.3 zeigt die neue Methodik. Anstatt die durch das NER-System erkannten Konzepte mittels der MeSH-Terme zu überprüfen und dadurch die Vielfalt der möglichen Konzepte von vornherein zu beschränken, wird nun ein anderer Ansatz zur Validierung der Konzepte gewählt. Die neue Methode sieht vor, mehrere biomedizinische Entity-Linker zur Extraktion von UMLS-Konzepten parallel auf PubMed-Abstracts durchzuführen und deren Ergebnisse in einem Ensemble zusammenzuführen. Wie in Unterabschnitt 3.1.5 vorgestellt wurde, gibt es in der Bioinformatik mehrere denkbare Entity-Linker zur Extraktion von UMLS-Konzepten. Da jedes dieser Modelle seine eigenen Methoden nutzt, ist zu erwarten, dass die Linker Unterschiede in den Ergebnissen liefern. Allerdings gibt es auch Überschneidungen. In diesem Umstand besteht die Chance, mittels Ensembles die Qualität der Markierungen zu erhöhen. Die zugrundeliegende Annahme ist, dass eine Konzeptmarkierung eher richtig ist,

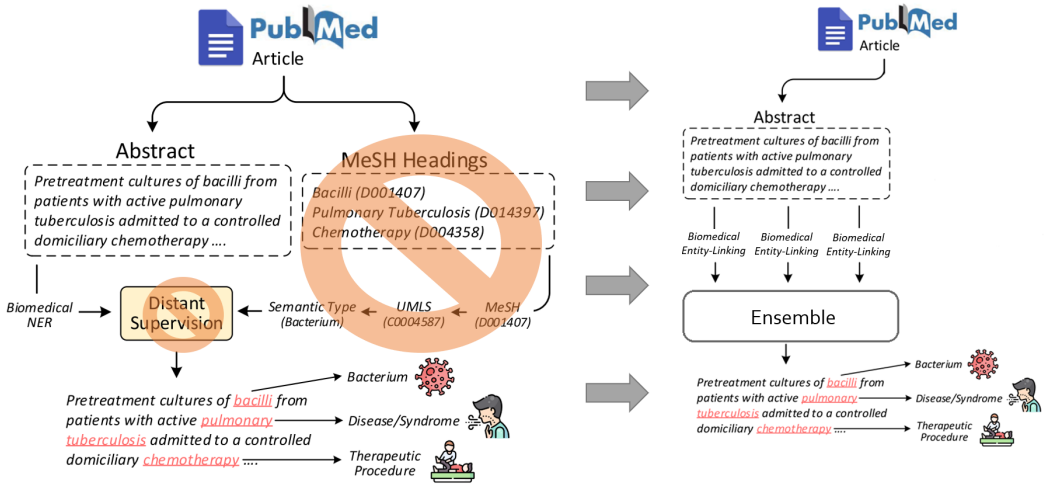

Abbildung 3.3: Die Abbildung zeigt die Methodik zur Erstellung der neuen Datensätze. Die Inspiration liefert die Methode zur Erstellung von PubMedDS auf der linken Seite. Anstatt des Distant Supervision mit den MeSH-Termen werden im neuen Ansatz mehrere biomedizinische Entity-Linker parallel durchgeführt und deren Ausgabewerte als Ensemble zusammengefügt.

je mehr Entity-Linker sie erkannt haben. Die Hoffnung ist es, dadurch Datensätze erstellen zu können, die im Vergleich zu PubMed eine Verbesserung im Recall liefern und gleichzeitig eine akzeptable Precision beibehalten. Im nächsten Unterabschnitt werden verschiedene Ensemble-Methoden vorgestellt, die das ermöglichen sollen. Das Ziel dieser Arbeit ist es, einen Korpus zu erstellen, der mehrere Datensätze beinhaltet. Dadurch soll zukünftigen Nutzern die Wahl überlassen werden welchen Datensatz sie für ihre Anwendung bevorzugen. Die verschiedenen Datensätze entstehen durch die Kombination aus ausgewählten Entity-Linkern und Ensemble-Methode.

3.2.2 Ensemble-Methoden

Für die drei in Unterabschnitt 3.1.5 vorgestellten Entity-Linker scispaCy *(s)*, cTAKES *(c)* und QuickUMLS *(q)* ergeben sich vier mögliche Kombinationen - *(s+c)*; *(s+q)*; *(q+c)*; *(s+q+c)* - um ein Ensemble zu formen. Für die Zusammenführung der Ausgabewerte der einzelnen Entity-Linker sind viele unterschiedliche Ensemble-Methoden denkbar. In dieser Arbeit werden drei vorgestellt. Wir definieren folgende Ensemble-Methoden:

- (1) *min1*: Es werden alle die Konzeptmarkierungen übernommen, die in mindestens einem der im Ensemble enthaltenen Entity-Linker vorhanden sind und exakt übereinstimmen.
- (2) *min2*: Es werden alle die Konzeptmarkierungen übernommen, die in mindestens zwei der im Ensemble enthaltenen Entity-Linker vorhanden sind und exakt übereinstimmen.
- (3) *all*: Es werden alle die Konzeptmarkierungen übernommen, die in allen im Ensemble enthaltenen Entity-Linkern exakt übereinstimmen.

Um ein besseres Verständnis für die drei definierten Ensemble-Methoden aufzubauen dient, das Beispiel aus Abbildung 3.4. Die erste Zeile zeigt den Ausgangstext des Beispiels. Auf diesen werden zunächst die drei Entity-Linker angewendet. Dadurch werden die Eigenschaften der Linker verdeutlicht. Bei allen drei Entity-Linkern ist es möglich, dass eine Textstelle mit mehr als einem Konzept verlinkt ist. Der Textstelle ‘cystic fibrosis‘ sind beispielsweise durch scispaCy mit C0010674 und C0010681 zwei Konzepte zugeordnet worden. Außerdem ermöglichen die Entity-Linker Überschneidungen bei den Konzeptmarkierungen. Zum Beispiel wurde durch cTAKES das Konzept C0033808 für ‘Pseudomonas‘ markiert, welches innerhalb der Grenzen der Konzeptmarkierung C0033809 liegt. Die nächsten drei Zeilen der Abbildung verdeutlichen nun die Ergebnisse der definierten Ensemble-Methoden für das *(s+q+c)*-Ensemble.

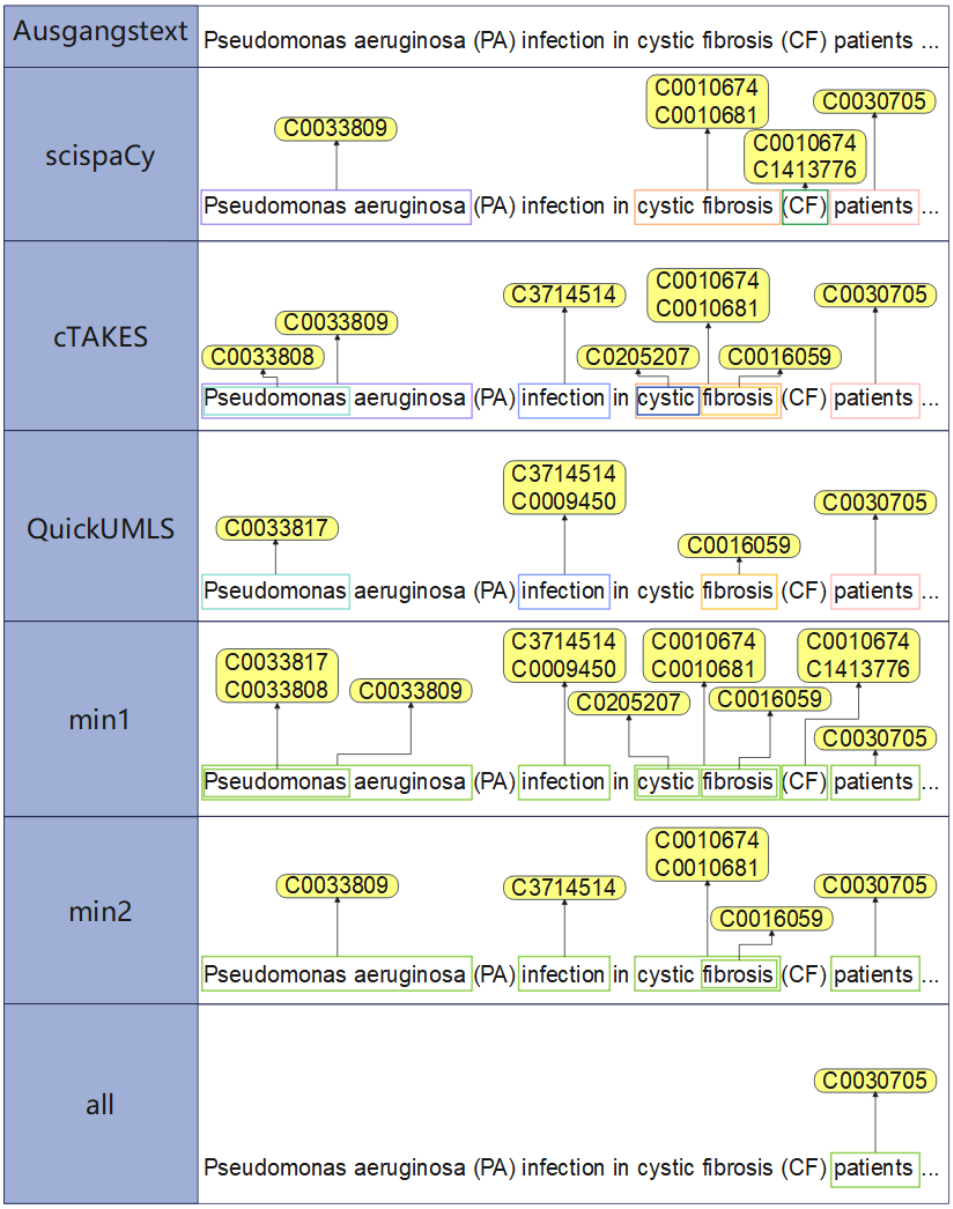

Abbildung 3.4: Beispielhafte Gegenüberstellung der Methoden zur Erstellung der Ensembles. Die ersten drei Felder der Grafik zeigen die Ausgabemarkierungen der drei Entity-Linker für den Beispieltext. Die letzten drei Felder zeigen die Ergebnisse für die drei vorgestellten Ensemble-Methoden. Die Ergebnisse gelten jeweils, wenn die Methode auf die Kombination aller drei Linker angewendet wird.

Für die *min1*-Methode beinhaltet das Ergebnis acht Markierungen mit insgesamt zwölf Konzepten. Im Vergleich liefert diese Methode die meisten Konzepte und Markierungen. Wenn einer Satzstelle von mindestens einem Linker ein Konzept zugeordnet wird, dann wird diese Satzstelle auch im Ergebnis der *min1*-Methode markiert. Die zugeordneten Konzepte sind dann alle Konzepte, die mindestens einmal dieser Satzstelle zugeordnet sind. Der Satzstelle 'Pseudonomas' werden also das Konzept C0033808 aus cTAKES und das Konzept C0033817 aus QuickUMLS zugeordnet.

Die *min2*-Methode beinhaltet fünf Markierungen und sechs Konzepte. Dabei werden nur Konzepte ins Ergebnis übernommen, die von mindestens zwei Linkern erkannt werden. Es ist egal, welche zwei Linker das Konzept beinhalten. Wichtig ist nur, dass bei einer Markierung die Konzepte der beiden Linker übereinstimmen. Das Konzept C3714514 ist beispielsweise sowohl von cTAKES als auch von QuickUMLS erkannt worden. Das Konzept C0033809 wiederum ist von cTAKES und scispaCy erkannt worden. Beide Konzepte werden ins Ergebnis übernommen. Der Satzteil 'Pseudomonas' wird sowohl von cTAKES als auch von QuickUMLS als eine Konzeptmarkierung erkannt. Allerdings erkennt cTAKES die Markierung als C033808 und QuickUMLS erkennt die Markierung als C0033817. Aus diesem Grund wird im Ergebnis der *min2*-Methode die Satzstelle nicht markiert, auch wenn zwei Linker dort fündig geworden sind.

Die *all*-Methode beinhaltet nur eine Markierung und ein Konzept. Dabei handelt es sich um den Satzteil 'patients'. Diesem wird das Konzept C0030705 zugeordnet, da dies im Beispielsatz die einzige Markierung ist, bei der alle drei Entity-Linker übereinstimmen. Logischerweise liefert diese Methode im Vergleich die wenigsten Konzepte und Markierungen, da im Ergebnis nur Konzepte übrig bleiben, die von allen Linkern gleich erkannt werden.

3.2.3 Technische Umsetzung

Nachfolgend geht es um die technische Realisierung der Methodik. Es geht darum, wie genau die drei Entity-Linker auf die Quelldaten angewendet werden können und vor allem, wie dabei eine große Datenmenge effektiv bearbeitet werden kann.

Ausgangspunkt

Als Quelldaten wird die PubMed-Baseline von 2021 verwendet.[13] Diese beinhaltet Daten zu 31.850.052 Dokumenten. Verwendet werden allerdings nur die Dokumente, für die der Titel und Abstract vorhanden sind. Diese Einschränkung führt zu einer Datenmenge von insgesamt 21.168.134 Dokumenten. Um aus den in Unterabschnitt 3.2.2 definierten Ensemble-Methoden einen Datensatz zu erstellen, müssen zuerst die drei Entity-Linker separat auf die Quelldaten angewendet werden. Da es sich dabei um mehrere Millionen Texte handelt, ist eine parallelisierte Verarbeitung notwendig. Die gewählte Software dafür ist Apache Spark bzw. PySpark. Für die Erstellung der virtuellen Umgebungen wird Conda verwendet.

Benutzerdefinierte UMLS-Wörterbücher

Keiner der vorgestellten Entity-Linker verwendet von Haus aus das volle UMLS mit allen Konzepten. scispaCy verwendet nur Vokabulare der Level 0, 1, 2 und 9. cTAKES verwendet die beiden Vokabularen SNOMED CT und RxNORM. Nur QuickUMLS nutzt die UMLS-Installation, die der Nutzer auf seinem PC installiert hat. Für jeden der drei Entity-Linker besteht die Möglichkeit ein benutzerdefiniertes Wörterbuch mit allen gewünschten UMLS-Quellvokabularen einzubinden. Um die Datensätze so umfassend wie möglich zu erhalten, wird versucht die Zahl der miteinbezogenen Vokabulare möglichst groß zu halten. Leider trat bei der Erstellung des Wörterbuchs für cTAKES ein Fehler auf, der im Laufe dieser Arbeit nicht behoben werden konnte. Dieser sorgt dafür, dass in cTAKES nicht das volle UMLS verwendet werden kann. Tabelle 3.3 listet die Vokabulare auf, die nicht in das Wörterbuch übernommen werden konnten. Auffällig ist, dass in den Abkürzungen aller fehleranfälligen Vokabulare die Zeichen '_', '-' oder '.' vorkommen. Dies könnte den Fehler hervorrufen. Um den Fehler zu beheben, fehlt allerdings das Verständnis für den Ablauf innerhalb des cTAKES-Prozesses. Daher wurde entschieden die fehlerbehafteten Vokabulare auszuschließen. Um die Ergebnisse der Entity-Linker einheitlich zu erhalten, werden für die Erstellung der Wörterbücher für scispaCy und QuickUMLS ebenfalls alle Vokabulare der Liste aus Tabelle 3.3 ausgeschlossen. Die genutzte UMLS-Version dieser Arbeit ist UMLS 2022AA.

[13] Der Download der aktuellen PubMed-Baseline ist unter: https://pubmed.ncbi.nlm.nih.gov/download/#annual-baseline möglich. Abgerufen am 12.06.2023

Tabelle 3.3: Eine Liste der Vokabulare die nicht in cTAKES funktionieren.

Name	Abkürzung	N. Einträge
Diagnostic & Statistical Manual of Mental Disorders, Fifth Edition	DSM-5	881
HL7 Vocabulary Version 2.5	HL7V2.5	4.911
HL7 Vocabulary Version 3.0	HL7V3.0	9.079
International Classification of Functioning, Disability & Health	ICF-CY	1.666
LOINC German, Austria Edition	LNC-DE-AT	6.812
LOINC German, Germany Edition	LNC-DE-DE	12.014
LOINC Greek, Greece Edition	LNC-EL-GR	2.304
LOINC Spanish, Argentina Edition	LNC-ES-AR	38.255
LOINC Spanish, Spain Edition	LNC-ES-ES	58.818
LOINC Spanish, Mexico Edition	LNC-ES-MX	83.185
LOINC Estonian, Estonia Edition	LNC-ET-EE	31.473
LOINC French, Belgium Edition	LNC-FR-BE	46.589
LOINC French, Canada Edition	LNC-FR-CA	46.807
LOINC French, France Edition	LNC-FR-FR	53.283
LOINC Italian, Italy Edition	LNC-IT-IT	85.049
LOINC Korean, Korea Edition	LNC-KO-KR	26.893
LOINC Dutch, Netherlands Edition	LNC-NL-NL	57.849
LOINC Polish, Poland Edition	LNC-PL-PL	4.139
LOINC Portuguese, Brazil Edition	LNC-PT-BR	58.425
LOINC Russian, Russia Edition	LNC-RU-RU	58.347
LOINC Turkish, Turkey Edition	LNC-TR-TR	52.129
LOINC Chinese, China Edition	LNC-ZH-CN	85.050
Medication Reference Terminology	MED-RT	3.467
NANDA-I Taxonomy II	NANDA-I	3.690
American College of Cardiology/American Heart Association	NCI_ACC-AHA	713
Clinical Data Interchange Standards Consortium	NCI_CDISC	24.468
CDISC Glossary Terminology	NCI_CDISC-GLOSS	767
Cancer Therapy Evaluation Program - SDC	NCI_CTEP-SDC	373
European Direcorate for the Quality of Medicines & Healthcare	NCI_EDQM-HC	1.026
NCI Dictionary of Cancer Terms	NCI_NCI-GLOSS	5.493
NCI Hugo Gene Nomenclature	NCI_NCI-HGNC	5.343
NCI Health Level 7	NCI_NCI-HL7	129
Prostate Imaging Reporting and Data System	NCI_PI-RADS	39

Konfiguration der Entity-Linker

Nachfolgend wird für alle drei Entity-Linker erläutert, wie diese in der vorliegenden Arbeit eingesetzt wurden:

scisapCy: Verwendet wird scispaCy Version 0.5.1. Das benutzte Modell ist 'en_core_sci_md' mit der Einstellung 'resolve_abbreviations' als 'True'. Die Einstellung führt dazu, dass Abkürzungen vor dem Entity-Linking in ihre Langform gebracht werden. Das Modell muss mit der Methode 'spacy.load()' geladen werden. Um für scispaCy ein benutzerdefiniertes Wörterbuch zu erstellen, kann mit kleinen Anpassungen der Code von allenai[14] genutzt werden. Anschließend muss ein benutzerdefinierter Linker erstellt werden, der das Wörterbuch nutzt. Der benutzerdefinierte Linker kann mit der Methode 'register_scispacy_linker()' erstellt werden. Danach muss er mit 'add_pipe()' zum 'en_core_sci_md'-Modell hinzugefügt werden. Zum Schluss des scispaCy Prozesses werden die Linking-Ergebnisse so gefiltert, dass für jede Textstelle nur die UMLS-Konzepte mit dem höchsten Score, übernommen werden.

[14]Der Code der verwendet wurde um das benutzerdefinierte Wörterbuch für scispaCy zu erstellen ist eine Adaption des Codes von allenai unter https://github.com/allenai/scispacy/blob/main/scripts/export_umls_json.py Abgerufen am 12.06.2023

cTAKES: Verwendet wird cTAKES Version 4.0.0.1. Um cTAKES für den vorliegenden Anwendungsfall nutzbar zu machen, muss ein eigener 'CollectionReader' und ein 'CollectionWriter' implementiert werden. Der 'CollectionReader' ist eine Erweiterung der Klasse 'JCasCollectionReader_ImplBase' und ermöglicht es der Pipeline das Format der Quelldaten zu lesen. Der 'CollectionWriter' ist eine Erweiterung der Klasse 'JCasConsumer_ImplBase' und sorgt dafür, dass die Linking-Ergebnisse im richtigen Format in eine Ausgabedatei geschrieben werden. Auf das UMLS wird in cTAKES über ein Wörterbuch-Lookup zugegriffen. Dafür wird im Normalfall die Datei 'sno_rx16_ab.xml' verwendet. Soll ein benutzerdefiniertes Wörterbuch verwendet werden, muss dies mittels der 'Dictionary Creator GUI' [15] erstellt werden. Als Ergebnis dieses Schrittes wird eine neue '.xml'-Datei erstellt, die der Pipeline hinzugefügt werden kann. Um das Entity-Linking durchführen zu können, müssen alle Pipeline-Komponenten in einer Piper-Datei zusammengefügt werden.

QuickUMLS: Verwendet wird QuickUMLS Version 1.4.1. Um QuickUMLS nutzen zu können, wird eine UMLS-Installation benötigt, die dann bei der Instanziierung mittels 'QuickUMLS(*PfadZuUMLS*)' angegeben wird. Es ist also nicht notwendig ein separates benutzerdefiniertes Wörterbuch zu erstellen. Die von QuickUMLS genutzten Vokabulare richten sich nach den Quellvokabularen, die in der UMLS-Installation ausgewählt wurden. Alle anderen Auswahlmöglichkeiten werden bei der Erstellung der QuickUMLS-Instanz bei ihrem Default-Wert belassen. Durch die 'match()'-Methode wird dann das Entity-Linking auf den gewünschten Text angewendet. Hier wurde die Variable 'best_match' auf 'True' gesetzt und die Variable 'ignore_syntax' auf 'False'.

Konfiguration von Spark

Genutzt wird Spark auf einem Cluster von 130 Nodes mit jeweils 80 GB nutzbarem Speicher. Um den Entity-Linking-Prozess für jeden Linker parallelisiert durchführen zu können, muss ein main-Programm mittels 'spark-submit' an das Cluster übergeben werden. Im 'spark-submit' enthalten ist das virtualenv, das main-Programm mit den Spark-Befehlen zum Parallelisieren, die Anzahl der Executors und die Speichergröße für den Driver und die Executors. Für **scispaCy** wurden 60 Executors mit jeweils 30 GB Speicher und ein Driver mit 4 GB Speicher genutzt. Die Bearbeitung hat 10 Stunden und 17 Mi-

[15] Eine Anleitung für das Erstellen eines benutzerdefinierten Wörterbuch für cTAKES ist unter https://cwiki.apache.org/confluence/display/CTAKES/Dictionary+Creator+GUI zu finden. Abgerufen am 12.06.2023

nuten gedauert. **cTAKES** benötigt 16 GB Speicher bei Executor und Driver und wurde mit 100 Executors ausgeführt. Die Rechenzeit von cTAKES betrug 8 Stunden und 55 Minuten. Bei **QuickUMLS** wurden 25 GB Speicher pro Executor und 2 GB für den Driver gewählt. QuickUMLS wurde mit 100 Executors durchgeführt und war nach 9 Stunden und 2 Minuten fertig.

Kapitel 4

Experimente

Das folgende Kapitel dient der Untersuchung der Ensemble-Methoden aus Unterabschnitt 3.2.2 und soll dabei helfen, den Nutzen des vorgestellten Korpus zu verdeutlichen. Außerdem werden in diesem Kapitel die formulierten Forschungsfragen beantwortet. Zunächst werden die Ensemble-Methoden mit Hilfe des MedMentions-Datensatzes evaluiert und verglichen. Danach wird für eine ausgewählte Methode in einer händischen Zweitevaluation untersucht, ob Verzerrungen der Precision-Werte vorliegen. Zum Schluss werden mittels deskriptiver Statistiken die Eigenschaften des Korpus verdeutlicht und vergleichbaren Datensätzen gegenübergestellt.

4.1 Evaluation durch MedMentions

Im folgenden Abschnitt wird die erste Forschungsfrage - 'Welche Kombinationen aus Ensemble-Methode und Entity-Linker-Auswahl führen zu den besten Ergebnissen?' - beantwortet. Dafür wird eine Evaluation der verschiedenen Kombinationsmöglichkeiten durchgeführt. Als Vergleichsdatensatz dient der in Unterabschnitt 3.1.3 vorgestellte MedMentions-Korpus.

4.1.1 Vorgehen

Als erstes werden die Ensemble-Methoden auf die Abstract-Texte der 4.392 Dokumente aus MedMentions angewendet. Dies erlaubt es, die Konzeptmarkierungen aller Ansätze mit den Konzeptmarkierungen des MedMentions-Datensatzes zu vergleichen und zu bewerten. Die dafür verwendeten Kennwerte sind die Precision (P) (vgl. Gleichung 4.1), der Recall (R) (vgl. Gleichung 4.2) sowie eine Auswahl an diversen Fβ-scores (vgl. Gleichung 4.3) [15].

Die Precision sagt im vorliegenden Anwendungsfall aus, wie groß der Anteil der korrekt vorhergesagten Konzepte an den insgesamt vorhergesagten Kon-

zepten ist. Die Formel für die Precision lautet:

$$\mathrm{P} = \frac{\sum \mathrm{TP}}{\sum \mathrm{TP} + \sum \mathrm{FP}} \tag{4.1}$$

Der Recall sagt aus, wie groß der Anteil der Konzepte des Vergleichsdatensatzes (MedMentions) ist, die richtig vorhergesagt wurden. Die Formel für den Recall lautet:

$$\mathrm{R} = \frac{\sum \mathrm{TP}}{\sum \mathrm{TP} + \sum \mathrm{FN}} \tag{4.2}$$

Der Fβ-score ist das gewichtete harmonische Mittel zwischen Precision und Recall. Für β kann ein beliebiger Wert eingesetzt werden. Dieser ist dann für die Gewichtung des Recalls zuständig. In der Formel ist der Recall also β-mal so wichtig wie die Precision. Die Formel für den Fβ-score lautet:

$$\mathrm{F}\beta = (1 + \beta^2) * \frac{\mathrm{P} * \mathrm{R}}{(\beta^2 * \mathrm{P}) + \mathrm{R}} \tag{4.3}$$

Abbildung 4.1 zeigt die Konfusionsmatrix für den vorliegenden Anwendungsfall. Die Matrix zeigt auf, in welchem Fall ein Konzept bei der Evaluation ein True Positive (TP), False Positive (FP), False Negative (FN) oder True Negative (TN) ist. Ein TP liegt vor, wenn für ein Konzept, welches im zu evaluierenden Datensatz markiert wurde, dieselbe Markierung in MedMentions existiert, also sowohl die Offsets als auch die CUI übereinstimmen. Besteht für ein Konzept, welches im Ensemble markiert wurde, keine übereinstimmende Konzeptmarkierung, handelt es sich um ein FP. Fälle, bei denen ein Konzept, das in MedMentions markiert ist, nicht im Ensemble markiert wurde, sind FN. Alle Wörter des Textes, die weder in MedMentions noch im Ensemble eine Konzeptmarkierung besitzen, zählen als TN. Diese Fälle sind allerdings für die Berechnung der Kennwerte irrelevant.

Vorhersage Konzept \ MedMentions Konzept	Positiv	Negativ
Positiv	TP	FP
Negativ	FN	TN

Abbildung 4.1: Konfusionsmatrix für den vorliegenden Vergleich.

Da die neuen Datensätze auch Textstellen besitzen, denen mehr als ein Konzept zugeordnet wurde, wird jedes dieser Konzepte separat evaluiert und einem der vier Fälle (TP, FP, FN, TN) zugeordnet. Mittels seqeval[1] werden durch ein Micro Averaging-Verfahren die Kennwerte für die Evaluation berechnet. Das bedeutet, dass zuerst die TP-, FP- und FN-Summen aller 4.392 Abstracts zusammengerechnet und anschließend mit den Gesamtsummen die Kennwerte berechnet werden.

4.1.2 Signifikanztest mit Bootstrapping

Um die Evaluationsergebnisse zuverlässig bewerten zu können, wird eine Untersuchung der Signifikanz durchgeführt. Dafür dient ein Signifikanztest auf Grundlage von Bootstrapping [10]. Untersucht wird, ob die Differenzen zwischen den Kennwerten der unterschiedlichen Methoden signifikant sind. Der Signifikanztest wird gemacht, um Vergleiche zwischen den Methoden durchführen zu können und die Erkenntnisse aus den Vergleichen statistisch zu untermauern. Für jeden Kennwert (Precision; Recall; F0.5-; F1-; F2-score) werden jeweils alle Methoden kreuzweise miteinander vergleichen. Daraus ergibt sich eine Gesamtmenge von:

$$78 \text{ Methoden Vergleiche} * 5 \text{ Kennwerte} = 390 \text{ Einzeltests} \tag{4.4}$$

Für jeden Einzeltest wird eine Bootstrappingstichprobe von 100.000 Fällen erstellt. Dafür werden zunächst die TPs, FPs und FNs der beiden zu vergleichenden Methoden auf zwei Listen zufällig neu verteilt. Die Voraussetzung dabei ist, dass die beiden Listen dieselbe Anzahl an Einträgen haben wie die Ursprungslisten der Methoden. Anschließend wird für jede der beiden neuen Listen der - durch den Test untersuchte - Kennwert neu berechnet und die Differenz gebildet. Die Differenzen werden in einer Häufigkeitsverteilung gesammelt. Nach 100.000 Wiederholungen kann durch den Erwartungswert und die Standardabweichung eine Normalverteilung erstellt werden. Diese wird genutzt, um die Signifikanz der tatsächlich beobachteten Differenz zu untersuchen. Ist die Wahrscheinlichkeit für die tatsächlich beobachtete Differenz oder einen extremeren Wert geringer als die Wahrscheinlichkeit für $\alpha/2$, dann ist die Beobachtung nicht signifikant. Ist das Gegenteil der Fall, dann ist die Beobachtung signifikant.

Da die Tests alle auf derselben Stichprobe durchgeführt werden, ist vor der Testentscheidung eine Korrektur des Signifikanzlevels α notwendig. Die Kor-

[1] Der Sourcecode von seqeval ist unter: https://github.com/chakki-works/seqeval zu finden. Abgerufen am 12.06.2023

rektur wird mittels der ungewichteten Bonferroni Prozedur durchgeführt [34]. Die Formel für α^{neu} lautet:

$$\alpha^{neu} = \alpha^{alt}/n \tag{4.5}$$

Dabei steht n für die Anzahl der durchgeführten Tests. Für α^{alt} wurde ein Wert von 0.05 gewählt. Daraus folgt für den vorliegenden Fall:

$$\alpha^{neu} = 0.05/390 \approx 0,00013 \tag{4.6}$$

4.1.3 Ergebnisse

Tabelle 4.1 beinhaltet die Ergebnisse der Evaluation. Die Tabelle bildet für alle denkbaren Datensätze die erreichten Kennwerte ab. Die Bezeichnung eines Datensatzes setzt sich aus der Ensemble-Methode und den verwendeten Entity-Linkern zusammen. Der Name *min1(s+q+c)* steht also beispielsweise für den Datensatz, der erstellt wurde durch die Ensemble-Methode *min1* und die Entity-Linker scispaCy (s), QuickUMLS (q) und cTAKES (c). Bei den ausgewerteten Kennwerten handelt es sich um Precision, Recall, F0.5-, F1- und F2-score. Um einen direkten Vergleich zum PubMedDS-Datensatz zu ermöglichen, sind auch die Kennwerte für PubMedDS enthalten. Außerdem sind Informationen über die Ergebnisse der Signifikanztests beinhaltet. Steht im Superskript eines Evaluationsergebnisses ein Buchstabe, ist das ein Zeichen für fehlende Signifikanz. Die Differenz zwischen dem betrachteten Evaluationsergebnis und dem Evaluationsergebnis, welches zum Buchstaben aus dem Superskript gehört, ist dann nicht signifikant. Ist ein Vergleich zwischen zwei Kennwerten nicht signifikant, dann sollte dies bei der Bewertung und Interpretation der Ergebnisse beachtet werden.

PubMedDS: Wie bereits in Unterabschnitt 3.1.4 beschrieben wurde, besitzt der PubMedDS-Datensatz eine hohe Precision von 0,903. Dem steht aber ein niedriger Recall von 0,053 gegenüber. Der niedrige Recall führt dazu, dass auch die Fβ-Werte mit 0,215 für F0.5, 0,100 für F1 und 0,065 für F2 vergleichsweise niedrig sind. Da der PubMedDS-Datensatz als Inspiration dieser Arbeit dient, bilden die Werte eine wichtige Vergleichsgrundlage für die Bewertung der Ensemble-Methoden.

Vergleich der Entity-Linker: Ein lohnender Vergleich ergibt sich aus den Ergebnissen der drei Entity-Linker, wenn diese, ohne Teil eines Ensembles zu sein, für sich stehen. Die Kennwerte zeigen, dass scispaCy und cTAKES ähnliche Entity-Linking-Ergebnisse aufweisen und QuickUMLS deutlich schlechtere Werte liefert. Während scispaCy und cTAKES jeweils eine Precision von

Tabelle 4.1: Verschiedene Kennwerte für PubMedDS, die drei Entity-Linker und alle neun Möglichkeiten der Ensembles. Als Goldstandard wurde der MedMentions-Datensatz verwendet. Des weiteren beinhaltet die Tabelle Ergebnisse der Signifikanztests. Steht nach einem Kennwert ein Buchstabe im Superskript, bedeutet dies, dass die Differenz zwischen dem betrachteten Wert und dem Wert des zum Buchstaben passenden Ansatzes nicht signifikant ist. In den Klammern stehen die Entity-Linker, die im Ensemble beinhaltet sind. s=ScispaCy, q=QuickUMLS, c=cTAKES.

	MedMentions				
	P	R	F0.5 score	F1 score	F2 score
a PubMedDS[2]	0,903	0,053	0,215	0,100	0,065
b ScispaCy	0,291 ^{d}	0,484 ^{d}	0,316 ^{d}	0,363 ^{d}	0,427 ^{d}
c QuickUMLS	0,162	0,139	0,157	0,150	0,143
d cTAKES	0,287 ^{b}	0,479 ^{b}	0,312 ^{b}	0,359 ^{b}	0,422 ^{b,i}
e min1(s+q+c)	0,197	0,585 ^{g}	0,227	0,294	0,420 ^{d}
f min1(s+q)	0,229 ^{h}	0,508	0,257 ^{g,h}	0,315 ^{h}	0,408 ^{h,i}
g min1(s+c)	0,224	0,579 ^{e}	0,255 ^{f,h}	0,323	0,440
h min1(q+c)	0,229 ^{f}	0,500	0,257 ^{f,g}	0,314 ^{f}	0,404 ^{f,i}
i min2(s+q+c)	0,399	0,411	0,401	0,405	0,409 ^{f,h}
j all(s+q+c)	0,528	0,098	0,281	0,166	0,117
k all(s+q)	0,370	0,112	0,253	0,172 ^{m}	0,130
l all(s+c)	0,501	0,377	0,470	0,430	0,397
m all(q+c)	0,344	0,118	0,249	0,175 ^{k}	0,136

ca. 0,29 und einen Recall von ca. 0,48 besitzen, liegen die Ergebnisse von QuickUMLS bei 0,162 für die Precision und 0,139 für den Recall. Aufgrund dieser Analyse sollte die Wahl von QuickUMLS als Teil des Ansatzes überdacht werden. Bei einem direkten Vergleich zwischen scispaCy und cTAKES ist scispaCy auf den ersten Blick bei jedem Kennwert um 0,004 bis 0,005 Punkte besser als cTAKES. Die Ergebnisse der Signifikanztests zeigen aber, dass diese Beobachtung nicht signifikant ist. Daher kann keine zuverlässige Aussage über einen Gewinner dieses Vergleichs gezogen werden. Eher unterstreicht der Signifikanztest die Annahme, dass die Qualität beider Entity-Linker auf demselben Level ist.

Vergleich der Ensemble-Methoden miteinander: Als nächstes werden die verschiedenen Methoden *min1*, *min2* und *all* miteinander verglichen. Auf den ersten Blick wird deutlich, dass die Methode *all*, unabhängig von den gewählten Entity-Linkern, tendenziell zu einer Erhöhung der Precision und die Methode *min1* zu einer Erhöhung des Recalls führen. Diese Beobachtung ist

[2]Die PubmedDS-Werte sind die Angaben von Vashishth et al. [36].

logisch, da bei *all* nur die Konzeptmarkierungen übernommen werden, die in allen Entity-Linkern vorhanden und daher am sichersten richtig sind. Das erhöht die Precision, senkt aber gleichzeitig den Recall. Auf der anderen Seite werden bei *min1* alle durch die Entity-Linker denkbaren Konzeptmarkierungen übernommen. Dies wirkt sich positiv auf den Recall aus, führt aber auch dazu, dass viele Konzeptmarkierungen übernommen werden, die nicht richtig sind. Dadurch sinkt bei *min1* gleichzeitig die Precision. Die Methode *min2* liefert in diesem Vergleich der drei Ensemble-Verfahren einen Mittelweg. Da für jede Konzeptmarkierung überprüft wird, ob sie durch einen zweiten Entity-Linker gefunden wurde, werden viele falsche Konzeptmarkierungen im Vergleich zu *min1* nicht übernommen. Dadurch steigt die Precision im Vergleich zu *min1*. Außerdem kann der Recall im Vergleich zu *all* erhöht werden. In *all* werden alle Konzeptmarkierungen verworfen, die nicht von allen Entity-Linkern erkannt werden. In *min2* werden aber nur Konzeptmarkierungen verworfen, die nicht von mindestens zwei Linkern erkannt werden. Dadurch fallen weniger Konzeptmarkierungen aus der Ergebnismenge heraus. Das führt im Vergleich zu einem höheren Recall. Die Methode *min2* ermöglicht es also, sowohl die Precision als auch den Recall zu erhöhen. Allerdings ist die Methode in keinem Kennwert die beste Methode.

Vergleich der Ensemble-Methoden von innen: Nun wird untersucht, welche Zusammensetzungen von Entity-Linkern am besten für die einzelnen Ensemble-Methoden funktionieren.

Die erste Methode ist *min1*. Diese eignet sich vor allem zur Erhöhung des Recalls. Der höchste Recall von 0,585 wird erreicht, wenn für *min1(s+q+c)* alle drei Entity-Linker in das Ensemble aufgenommen werden. Dies ist sowohl innerhalb des Ansatzes als auch beim Gesamtvergleich allen anderen Methoden der höchste Recall. Das Ergebnis des Signifikanztests zeigt allerdings, dass der Unterschied zu *min1(s+c)* und dessen Recall von 0,579 nicht signifikant ist. Da *min1(s+c)* zusätzlich in allen anderen vier Kennwerten signifikant bessere Ergebnisse liefert und sogar den höchsten F2-score besitzt, empfiehlt es sich, bei einer Priorisierung des Recalls den *min1(s+c)*-Datensatz zu wählen. Die beiden anderen Ensembles des *min1*-Ansatzes sind *min1(s+q)* und *min1(q+c)*. Sowohl *min1(s+q)* mit einer Precision von 0,229 und einem Recall von 0,508 als auch *min1(q+c)* mit einer Precision von 0,229 und einem Recall von 0,500 sind im Vergleich mit den anderen beiden Ansätzen keine guten Alternativen.

Die nächste Methode ist *min2*. Wie bereits beschrieben, ist diese ein Mittelweg zwischen der Optimierung der Precision und der Optimierung des Recalls. Die Methode setzt voraus, dass mindestens drei Entity-Linker im Ensemble

beinhaltet sind. Das führt dazu, dass es mit *min2(s+q+c)* nur eine Möglichkeit gibt, ein Ensemble zu bilden. Dieses besitzt eine Precision von 0,399, einen Recall von 0,411, einen F0.5-score von 0,401, einen F1-score von 0,405 und einen F2-score von 0,409. Der Ansatz liefert zwar für keinen Kennwert das höchste Ergebnis, allerdings ist er immer unter den besten Ansätzen dabei. Der F0.5-score und der F1-score habend jeweils die zweithöchsten Werte aller Ansätze und liegen 0,069 bzw. 0,025 Punkte unter den Werten von *all(s+c)*. Der F2-score bietet mit 0,409 ebenfalls ein gutes Ergebnis und liegt gerade einmal 0,031 Punkte unter dem bestplatzierten Ansatz *min1(s+c)*. Für den *min2*-Ansatz lässt sich festhalten, dass er vor allem durch seine Ausgeglichenheit zwischen Precision und Recall eine interessante Alternative bietet.

Die letzte Methode ist *all*. Diese eignet sich vor allem zur Erhöhung der Precision. Die höchste Precision von 0,528 wird erreicht, wenn alle drei Entity-Linker in das Ensemble aufgenommen werden, also bei *all(s+q+c)*. Dies ist sowohl innerhalb des Ansatzes als auch beim Gesamtvergleich mit den anderen Methoden die höchste Precision. Allerdings liefert diese Variante einen Recall von nur 0,098. Wird ein Datensatz mit einer möglichst hohen Precision benötigt, sollte daher die Wahl auf PubMedDS fallen. PubMedDS besitzt eine Precision von 0,903 und liegt damit um 0,375 Punkte höher als die Precision von *all(s+q+c)*. Die zweitbeste Precision wird durch das Ensemble *all(s+c)* mit einem Wert von 0,501 erreicht. Zusätzlich zu der hohen Precision besitzt die Variante einen Wert von 0,377 beim Recall. Das führt dazu, dass sowohl der F0.5-score mit 0,470 als auch der F1-score mit 0,430 die höchsten Werte aller Ansätze aufweisen. Da auch alle Signifikanztests positiv sind, ist *all(s+c)* die wohl beste Wahl, wenn ein Datensatz mit einem guten Mittelweg zwischen Precision und Recall gesucht wird. Die beiden anderen Ensembles der *all*-Methode sind *all(s+q)* und *all(q+c)*. Sowohl *all(s+q)* mit einer Precision von 0,370 und einem Recall von 0,112 als auch *all(q+c)* mit einer Precision von 0,344 und einem Recall von 0,118 sind im Vergleich mit den anderen beiden Varianten – wie bereits bei der *min1*-Methode - keine guten Alternativen.

Schlussfolgerung: Zusammenfassend lässt sich für die Beantwortung der ersten Forschungsfrage - 'Welche Kombinationen aus Ensemble-Methode und Entity-Linker-Auswahl führen zu den besten Ergebnissen?' - festhalten, dass die Entscheidung für einen Gewinneransatz von der persönlichen Priorisierung der Precision bzw. des Recalls abhängig gemacht werden sollte. Soll der Datensatz später für einen Anwendungsfall verwendet werden, bei der die Richtigkeit der einzelnen Markierungen wichtiger ist als deren pure Anzahl, sollte eher auf den Wert der Precision oder des F0.5-scores geachtet werden. Für den Fall

wäre PubMedDS oder die Methode *all(s+c)* der beste Ansatz. Soll der Datensatz allerdings für eine Methode verwendet werden, welche eine hohe Anzahl an Markierungen erfordert, sollte auf den Recall oder den F2-score geachtet werden. In dem Fall sollte die Wahl auf eine der Varianten *min1(s+q+c)* und *min1(s+c)* fallen. Ist sowohl die Precision als auch der Recall wichtig, dann sollte die Entscheidung auf Basis des F1-scores getroffen werden. Die beste Wahl wäre in diesem Fall wieder die Methode *all(s+c)*. Aufgrund der schlechten Ergebnisse von QuickUMLS ist unklar, welche Methoden die besten wären, wenn ein besserer dritter Entity-Linker gewählt worden wäre.

4.2 Händische Zweitevaluation

Bei der Erstellung von MedMentions und dem hier vorgestellten Datensatzkonzept liegen unterschiedliche Rahmenbedingungen vor. Es gibt Unterschiede in der Anzahl möglicher Konzepte pro markierter Textstelle, im Umgang mit Überlappungen, bei der Einschränkung semantischer Typen und aufgrund von Änderungen im UMLS. Diese Unterschiede in den Rahmenbedingungen führen unter Umständen zu verzerrten Evaluationsergebnissen. Um dies zu untersuchen, wird eine händische Zweitevaluation durchgeführt. Die Motivation zu dieser Zweitevaluation liegt in der Vermutung, dass eine Großzahl der als FP markierten Konzepte nur aufgrund der unterschiedlichen Rahmenbedingungen als FP markiert wurden. Wenn dies der Fall wäre, könnte das eine Verzerrung der in Unterabschnitt 4.1.3 ermittelten Precision-Werte zufolge haben. Nachfolgend werden die Unterschiede in den Rahmenbedingungen genauer beschrieben. Anschließend wird das Vorgehen bei der händischen Zweitevaluation erläutert. Zum Schluss werden die Evaluationsergebnisse vorgestellt und interpretiert. Dadurch soll die zweite Forschungsfrage - 'Sind die Precision-Werte der Evaluation aufgrund von ungleichen Bedingungen bei der Erstellung der Datensätze verfälscht?' - beantwortet werden.

4.2.1 Unterschiede zu MedMentions

Nachfolgend werden die Unterschiede zwischen MedMentions und dem hier vorgestellten Korpus erläutert, aufgrund derer eine händische Zweitevaluation sinnvoll ist.

1. Der erste Unterschied besteht in der Anzahl möglicher Konzepte pro identifizierter Textstelle. MedMentions beinhaltet genau ein Konzept pro Markierung. Es ist daher denkbar, dass bei der Erstellung der MedMentions-Markierungen häufig Situationen entstanden sind, bei denen sich die Experten zwischen mehreren potenziellen Konzepten ent-

scheiden mussten. Die Ensemble-Methoden dieser Arbeit erlauben es andererseits, mehrere Konzepte pro Textstelle zu markieren. Wenn nun in einer Markierung mehr als ein Konzept gekennzeichnet ist, kann höchstens eins davon mit dem Konzept aus MedMentions übereinstimmen. Daher ist es in diesen Fällen mit der ursprünglichen Evaluation aus Abschnitt 4.1 maximal möglich, die Summe der TPs um eins zu erhöhen. Die übriggebliebenen Konzepte der Textstelle gehen dann alle als FP in die Evaluation ein. Allerdings enthält das UMLS auch Konzepte, die sich untereinander sehr stark ähneln. Dadurch ist eine Konzeptmarkierung oft nicht, wie durch MedMentions vorausgesetzt, eindeutig. In Fällen, in denen einer Textstelle mehr als ein Konzept zugeordnet ist, ist es daher sinnvoll, händisch zu überprüfen, ob die als FP gesetzten Konzepte evtl. ebenfalls richtig, also ein TP, sein könnten.

2. Eine weitere Differenz besteht darin, dass MedMentions keine Überlappungen der Markierungen erlaubt, während dies bei den hier vorgestellten Methoden möglich ist. Beispielsweise ist im *min1(s+q+c)*-Datensatz für die Textstelle 'Pseudomonas aeruginosa' sowohl die Sequenz 'Pseudomonas aeruginosa' als auch die Sequenz 'Pseudomonas' mit einem UMLS-Konzept markiert. In der Evaluation ist nun aufgrund der Überlappung der Markierungsgrenzen von vornherein ausgeschlossen das beide Konzepte als TP gewertet werden, obwohl das Potenzial gegeben ist, dass beide Werte TP sein könnten.

3. Bei der Erstellung des MedMentions-Korpus sind nur Konzepte in das Ergebnis übernommen worden, welche mit mindestens einem von 21 ausgewählten semantischen Typen oder einem untergeordnetem Elementen dieser Typen verbunden sind. Die genauen 21 Typen sind bereits in Unterabschnitt 3.1.3 beschrieben und können in Tabelle 3.1 nachgelesen werden. Aus dieser Eingrenzung folgt, dass es UMLS-Konzepte gibt, welche durch eine der Methoden dieser Arbeit markiert werden, aber nicht in den möglichen UMLS-Konzepten aus MedMentions vorhanden sind. Ein Beispiel sind Konzepte des semantischen Typen 'Clinical Drug'. Diese Fälle sind in der ersten Evaluation immer FP. Es besteht aber die Möglichkeit, dass sie TP wären, wenn der semantische Typ, welchem sie angehören, in MedMentions erlaubt wäre.

4. Das UMLS und seine Quellvokabulare entwickeln sich ständig weiter. Es gibt in einem neuen Release immer wieder Konzepte die wegfallen, neu gebildet werden oder ihre CUI ändern. Zum Zeitpunkt der Erstellung von MedMentions (UMLS 2017AA) gab es 3.465.486 Konzepte im UMLS. Im 2022AA Release waren es 4.553.796 und im aktuellen UMLS

2023AA sind es nur noch 3.313.382 Konzepte.[3] Diese Zahlen belegen die stetige Veränderung des UMLS über die Zeit hinweg. Das führt dazu, dass sich Konzeptmarkierungen aus neueren Ansätzen und von MedMentions schon aufgrund der Veränderung des UMLS unterscheiden können und daher als FP markiert werden, obwohl sich beispielsweise nur die CUI geändert hat.

4.2.2 Vorgehen

Aufgrund des großen Zeitaufwands, den eine händische Evaluation mit sich bringt, können nicht alle Kombinationen aus Ensemble-Methode und ausgewählten Entity-Linkern untersucht werden. Diese Arbeit beschränkt sich daher auf die händische Zweitevaluation eines einzigen Ansatzes. Dabei handelt es sich um den *all(s+c)*-Datensatz, also denjenigen, der in Unterabschnitt 4.1.3 den höchsten F1-score erreicht hat. Außerdem werden nicht alle Texte aus MedMentions neu evaluiert, sondern nur eine zufällige Auswahl von 100 Texten. Bei der Evaluation werden alle in den Texten gesetzten Markierung nacheinander betrachtet und nach TP und FP neu bewertet. Nicht markierte Textstellen werden nicht überprüft. Um für die Bewertung der Entitäten ein einheitliches und nachvollziehbares Vorgehen zu garantieren, wird vorher ein Regelwerk definiert, welches vorgibt, wann ein Konzept als TP bewertet werden soll. Trifft für ein Konzept keine der Regeln zu, wird es als FP bewertet. Ist keine eindeutige Zuordnung anhand der Regeln möglich, wird das Konzept ebenfalls als FP bewertet.

Abbildung 4.2 stellt das Regelwerk in Form eines Diagramms, ähnlich einem Entscheidungsbaum, dar. Für jedes UMLS-Konzept, dass in *all(s+c)* markiert ist, wird der Diagrammfluss durchlaufen und dadurch entschieden, ob das Konzept als TP oder FP markiert wird. Das Diagramm besteht aus fünf Swimlanes. Diese stellen die fünf Grundfälle für Mögliche TP-Markierungen dar.

Der **erste Fall** tritt ein, wenn für das Konzept, das in *all(s+c)* markiert ist, dasselbe Konzept an derselben Stelle in MedMentions auch markiert ist. Markierungen dieser Gruppe gehen - genau wie in der ursprünglichen Evaluation – immer als TPs in die Evaluation ein.

[3] Aktuelle Zahlen des UMLS 2023AA sind unter https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html zu finden. Die historischen Zahlen sind im Archiv unter https://www.nlm.nih.gov/research/umls/archive/archive_home.html zu finden. Jeweils abgerufen am 12.06.2023

Beim **zweiten Fall** ist in *all(s+c)* und MedMentions zwar dieselbe Textstelle markiert, allerdings sind unterschiedliche UMLS-Konzepte verlinkt. In der Evaluation aus Abschnitt 4.1 werden diese Markierungen immer als FP gezählt. In der händischen Evaluation gibt es nun aber drei Möglichkeiten, dass das Konzept noch in einen TP geändert wird. Die erste Möglichkeit besteht, wenn das Konzept aus *all(s+c)* eine allgemeinere Form des Konzeptes aus MedMentions ist. Ein Beispiel hierfür wären die Konzepte 'Height' und 'Body Height'. Ist in *all(s+c)* für eine Textstelle das Konzept 'Height' und in MedMentions für dieselbe Textstelle das Konzept 'Body Height' markiert, dann ist das *all(s+c)*-Konzept ursprünglich als FP bewertet. Da 'Height' aber einfach ein allgemeinerer Begriff für 'Body Height' ist, wird die Markierung durch diese Regel nun als TP bewertet. Bei der zweiten Möglichkeit handelt es sich um die Umkehrung der ersten. Sie tritt ein, wenn das Konzept aus *all(s+c)* eine spezifischere Form des Konzeptes aus MedMentions ist. Allerdings muss bei der Bewertung darauf geachtet werden, ob die Spezifizierung des Konzeptes auch Sinn ergibt. Ein passendes Beispiel ist, wenn für dieselbe Textstelle die Konzepte 'Scientific Study' in *all(s+c)* und 'Study' in MedMentions markiert sind. 'Scientific Study' ist ein spezifischerer Begriff für 'Study'. Handelt es sich bei der markierten Textstelle tatsächlich um eine wissenschaftliche Studie und nicht nur eine einfache Studie, dann erlaubt es das Regelwerk, diese als TP zu markieren. Die letzte Möglichkeit liegt bei Konzepten vor, deren CUI sich im Laufe der Zeit geändert haben. Ist die CUI des *all(s+c)*-Konzepts die neuere Version der CUI des MedMentions-Konzepts, dann kann die Markierung ebenfalls als TP bewertet werden. Das Konzept 'Cystic Echinococcosis' beispielsweise besaß zum Zeitpunkt der Erstellung von MedMentions die CUI C4303092. Mittlerweile ist die zugehörige CUI aber die C4553297. Wurde eine Textstelle in *all(s+c)* mit der neuen CUI C4553297 markiert und dieselbe Textstelle in MedMentions mit der alten CUI C4303092, sollte diese Markierung als TP gezählt werden.

Der **dritte Fall** tritt ein, wenn die Textstelle des *all(s+c)*-Konzepts innerhalb einer in MedMentions markierten Textstelle liegt. Ursprünglich ist dieser Fall aufgrund der Überlappung als FP markiert. Die Regel besagt nun aber, dass eine Textstelle dann doch als TP markiert wird, wenn das *all(s+c)*-Konzept einen Teil des MedMentions-Konzepts darstellt. Ein Beispiel für das Eintreten dieser Regel ist der Text 'A review of the literature revealed a total of 30 patients with SLE - AIMF reported to - date.' In diesem Text wurde in *all(s+c)* das Konzept 'Literature' für die Textstelle 'literature' markiert. In MedMentions wurde für denselben Text das Konzept 'Review Literature' für die Textstelle 'review of the literature' markiert. Eigentlich wäre diese Konzeptmarkierung als FP in die Evaluation eingegangen. Da 'Literature' aber

ein Teilbegriff von 'Review Literatur' ist, sollte die Textstelle als TP bewertet werden.

Der **vierte Fall** ist das Gegenteil des dritten Falls. Und zwar tritt dieser ein, wenn die Textstelle des in *all(s+c)* markierten Konzepts eine oder mehrere in MedMentions markierte Textstellen umschließt. Auch hier handelt es sich in der ersten Instanz immer um ein FP. Damit aus dieser Konzeptmarkierung ein TP-Fall gemacht werden darf, besagt die Regel, dass das umschlossene MedMentions-Konzept ein Teil des *all(s+c)*-Konzeptes sein muss. Ein Beispiel betrifft den Text 'We identified an increase in poor nutrition in surgical patients.' In diesem Text wurde in *all(s+c)* das Konzept 'Surgical Patients' für die Textstelle 'surgical patients' markiert. In MedMentions wurden für denselben Text die zwei Konzepte 'Operative Surgical Procedures' für die Textstelle 'surgical' und 'Patients' für die Textstelle 'patients' markiert. Die Markierung aus *all(s+c)* setzt sich also aus den zwei Markierungen aus MedMentions zusammen und kann daher als TP bewertet werden, obwohl sie ursprünglich als FP bewertet wurde.

Der **fünfte Fall** tritt ein, wenn die Textstelle des in *all(s+c)* markierten Konzepts nicht in MedMentions als Konzept markiert wurde und es auch keine Überschneidungen zu in MedMentions markierten Konzepten gibt. Logischerweise werden diese Markierungen als FP bewertet. Allerdings gibt es nun zwei Ausnahmen, bei denen eine Änderung zu einem TP überprüft wird. Die erste Ausnahme sind Konzepte, die einem der semantischen Typen angehören, welche nicht in MedMentions vorhanden sind. Trifft dies zu und das markierte Konzept macht an der vorliegenden Textstelle Sinn, wird die Markierung als TP bewertet. Ein Beispiel hierfür ist der Text 'The enrichment of bovine culture medium with 1 μM crocetin reduced incidence of apoptosis.'. In *all(s+c)* wurde für diesen Text an der Textstelle 'incidence' das Konzept 'Incidence' markiert. Diesem ist der semantische Typ 'Quantitative Concept' mit der TUI T081 zugeordnet. 'Quantitative Concept' ist allerdings einer der semantischen Typen, deren Konzeptmarkierungen in MedMentions verworfen werden. Da die Markierung an der Textstelle Sinn ergibt, wird das Konzept nachträglich zu einem TP geändert. Die zweite Ausnahme liegt vor, wenn das *all(s+c)*-Konzept zwar nicht dieselbe Markierung in MedMentions besitzt, das Konzept aber an einer anderen Stelle in MedMentions bereits markiert war und die Konzeptmarkierung an der *all(s+c)*-Textstelle ebenfalls Sinn macht. Es sind also Fälle, in denen in MedMentions keine konsequente Markierung vorliegt. Ein Beispiel dafür ist das Konzept 'study'. Häufig fangen Abstracts mit 'This study ...' oder ähnlichen Formulierungen an. In MedMentions ist das Konzept 'study' zwar bei einigen Texten markiert, aber nicht bei jedem Auftreten. Es ist aber

davon auszugehen, dass das Konzept jedes Mal gesetzt werden könnte. Fängt daher ein Text mit 'This study ...' an und die Konzeptmarkierung für 'study' ist in *all(s+c)* gegeben aber in MedMentions nicht, dann sollte das Konzept dennoch als TP markiert werden.

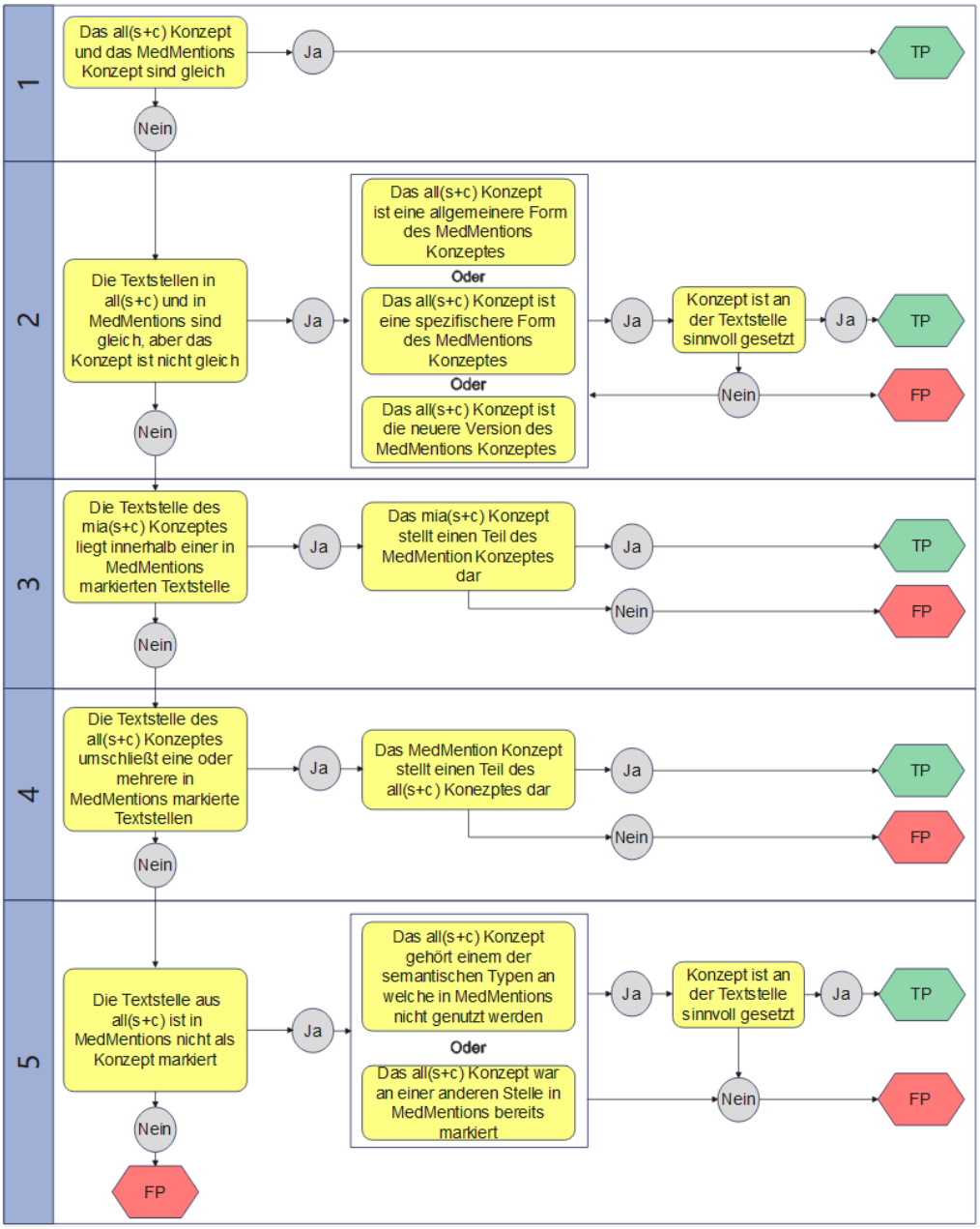

Abbildung 4.2: Das Regelwerk, anhand dessen die händische Evaluation durchgeführt wurde. Das Regelwerk kann wie ein Entscheidungsbaum durchlaufen werden.

4.2.3 Ergebnisse

Da zur händischen Zweitevaluation keine externen Mittel zur Verfügung stehen, wird diese eigenhändig durchgeführt. Dadurch kann eine unvoreingenommene und fachlich einwandfreie Bewertung nicht garantiert werden. Die klaren Regeln wirken diesem Umstand entgegen. Nichtsdestotrotz sollten die nachfolgenden Evaluationsergebnisse nur unter Vorbehalt betrachtet werden.

Um eine zuverlässige Aussage zu treffen, werden zunächst die Precision und der Recall für die 100 zufällig ausgewählten Abstracts mit der ursprünglichen Evaluation neu berechnet. Für die 100 Testfälle liegen die Precision bei 0,480 und der Recall bei 0,341 und damit etwas unter den Werten der vollen Evaluation mit den 4.392 MedMentions-Abstracts (P=0,501; R=0,377). Nach der händischen Zweitevaluation liegt die Precision bei 0,623. Dies entspricht einem Unterschied von 14,3% bei den Precision-Werten. Über den Wert des Recalls kann keine Aussage getroffen werden, da sich der Recall auf den Vergleichsdatensatz (hier MedMentions) bezieht. Der Recall sagt aus, wie hoch der Anteil der aus dem Vergleichsdatensatz gefunden Entitäten ist. Da sich der Vergleichsdatensatz nicht verändert bzw. nicht angezweifelt werden soll, macht es keinen Sinn, eine Anpassung beim Recall durchzuführen. Die Precision hingegen ist aus Sicht des eigenen Datensatzes und beschreibt, wie hoch der Anteil der korrekten Markierungen ist. Daher macht es Sinn zu untersuchen, ob der Precision-Wert höher sein könnte für den Fall, dass andere Bewertungskriterien gewählt werden. Zur Beantwortung der zweiten Forschungsfrage - 'Sind die Precision-Werte der Evaluation aufgrund von ungleichen Bedingungen bei der Erstellung der Datensätze verfälscht?' - lässt sich festhalten, dass der Precision-Wert des Ansatzes *all(s+c)* bei der Evaluation mit MedMentions zu niedrig ist und von einer erhöhten Precision von etwa +14% ausgegangen werden kann.

4.3 Deskriptive Statistiken

Der nachfolgende Abschnitt dient dazu, die Datensätze des Korpus genauer zu beschreiben und dadurch greifbarer zu machen. Dafür werden deskriptive Statistiken berechnet und ein Vergleich zu MedMentions und PubMedDS gezogen. Außerdem wird die dritte Forschungsfrage - 'Welchen Mehrwert liefert der vorgestellte Korpus im Vergleich zu anderen Datensätzen?' - untersucht. Diese wird in zwei Schritten beantwortet. Als erstes werden allgemeine Kennzahlen aufgelistet und verglichen. Anschließend werden die Konzeptmarkierungen genauer betrachtet. Dafür werden die semantischen Typen in allgemeinere Gruppen aufgeteilt und anschließend verglichen wie Häufig diese Gruppen in

Tabelle 4.2: Deskriptive Statistiken zum Vergleich der zwei ausgewählten Datensätze *min1(s+q+c)* und *all(s+c)* mit MedMentions (MM) und PubMedDS (PMDS).

	min1(s+q+c)	all(s+c)	MM	PMDS
Dokumente	21.168.134	21.168.134	4.392	13.197.430
einzigartiger UMLS-Konzepte	803.483	517.508	34.724	44.881
markierte Textstellen	2.642.771.731	970.964.807	352.496	57.943.354
UMLS-Konzepte	4.466.216.545	1.101.206.434	352.496	159.795.085
markierte Textstellen/Dokument	124,8	45,9	80,3	4,4
Konzepte/Dokument	211,0	52,0	80,3	12,1
Sätze	200.384.380	200.384.380	42.602	127.670.590
Token	4.792.379.324	4.792.379.324	1.176.058	3.019.566.668
annotierte Token	2.101.743.035	1.060.279.978	579.839	67.148.729
Anteil der annotierten Token	43,9%	22,1%	49,3%	2,2%
Token/markierter Textstellen	1,2	1,2	1,6	1,2
Token/Dokument	226,4	226,4	267,8	228,8
annotierte Token/Dokument	99,3	50,1	132,0	5,1
Sätze/Dokument	9,5	9,5	9,7	9,7
Token/Satz	23,9	23,9	27,6	23,6

den Datensätzen vorkommen. Aus Gründen der Übersichtlichkeit wurden für die deskriptiven Statistiken zwei Datensätze des Korpus ausgewählt. Der erste ist der *min1(s+q+c)*-Datensatz. Dieser wurde ausgewählt, da er in der Evaluation aus Abschnitt 4.1 die höchsten F0.5- und F1-scores erreicht hat. Beim zweiten Datensatz handelt es sich um *all(s+c)* und damit um den, der den höchsten Recall aller vorgestellten Varianten besitzt. Die beiden ausgewählten Ansätze werden mit den beiden bestehenden Datensätzen MedMentions und PubMedDS verglichen.

4.3.1 Allgemeine Kennwerte

Für den ersten Teil der deskriptiven Statistiken werden allgemeine Kennwerte verglichen. Tabelle 4.2 stellt die Kennwerte der vier Datensätze gegenüber. Dadurch wird deutlich, welche neuen Qualitäten der entwickelte Korpus liefern kann.

Die Anzahl der Dokumente liegt bei den neuen Datensätzen mit 21.168.134 weit über dem Wert von MedMentions mit 4.392 Dokumenten und ist auch höher als der PubMedDS-Wert von 13.197.430. Die große Anzahl der Dokumente wirkt sich auch auf die Anzahl der markierten UMLS-Konzepte aus. Bei den beiden neuen Datensätzen sind in Summe deutlich mehr markierte

Textstellen und dadurch auch mehr UMLS-Konzepte als bei den Vergleichsdatensätzen vorhanden. Der Datensatz *min1(s+q+c)* besitzt zum Beispiel über 4 Mrd. Konzepte verteilt auf ca. 2,5 Mrd. markierte Textstellen. Das sind ca. 12.000-mal mehr UMLS-Konzepte als in MedMentions und ca. 28-mal die Anzahl der UMLS-Konzepte von PubMedDS. Bei der Betrachtung der durchschnittlich markierten Textstellen bzw. durchschnittlichen Konzepte pro Dokument ist dieser große Unterschied nicht mehr zu beobachten. Der Datensatz *min1(s+q+c)* besitzt hier mit 124,8 Textstellen bzw. 211,0 Konzepten zwar immer noch mehr als die anderen Datensätze, der Unterschied zu MedMentions (80,3 bei beiden Kennzahlen) ist aber logischerweise nicht mehr so groß wie bei den Gesamtzahlen. Der Datensatz *all(s+c)* liegt jetzt mit 45,9 Textstellen und 52,0 Konzepten sogar hinter MedMentions. Der Vergleich zu PubMedDS (4,4 markierte Textstellen und 12,1 Konzepte) zeigt allerdings immer noch, dass durch die neuen Datensätze eine umfangreichere Konzeptmarkierungsdichte bei größerer Gesamtdatensatzgröße vorhanden ist.

Durch Tokenisierung der Texte wurden Kennwerte bezüglich der Anzahl von Sätzen und Token ermittelt. Die Kennwerte zeigen, dass die neuen Datensätze aus ca. 200 Mio. Sätzen und ca. 4,8 Mrd. Token bestehen und damit in diesen Kennwerten größer sind als die Vergleichsdatensätze. Bei PubMedDS sind es ca. 127 Mio. Sätze und ca. 3 Mrd. Token. Bei MedMentions sind es dann nur noch ca. 42,5 Tsd. Sätze und 1,1 Mio. Token. Neben den Gesamtsummen sind auch die Durchschnittswerte der Sätze und Token angegeben. Die Texte, die in MedMentions beinhaltet sind, besitzen im Schnitt mit 9,7 bzw. 267,8 die meisten Sätze und Token. Ein Satz in MedMentions besteht durchschnittlich aus 27,6 Token. PubMedDS besitzt ebenfalls 9,7 Sätze pro Dokument. Allerdings sind die Sätze mit 23,6 Token pro Satz kürzer als die Sätze des MedMentions-Korpus. Die Werte der beiden Ensemble-Datensätze ähneln den Werten von PubMedDS. Sie liegen bei 9,5 Sätzen pro Dokument, 226,4 Token pro Dokument und 23,9 Token pro Satz. Diese Kennwerte zur Tokenisierung machen deutlich, dass die in MedMentions verwendeten Texte mehr und längere Sätze besitzen. Ein Grund dafür könnte darin liegen, dass in den automatisierten Verfahren dieser Arbeit und von PubMedDS keine Kriterien für die Auswahl von Texten existieren. Es wird lediglich für jeden Abstract überprüft, dass der Text nicht leer ist. Dadurch ist denkbar, dass einige sehr kurze Abstracts in die Datensätze übernommen wurden. Im Gegensatz dazu hatten die Autoren von MedMentions die Möglichkeit, Texte auszusortieren, die nicht ihren Anforderungen entsprachen.

Besonders interessant sind die Kennwerte 'Anteil der annotierten Token' und 'Token / markierter Textstelle'. Hier hebt sich MedMentions mit einem

Anteil von 49,3 % an annotierten Token und einer durchschnittlichen Länge von 1,6 Token pro markierter Textstelle deutlich von den anderen Datensätzen ab. Während *min1(s+q+c)* noch 43,9% annotierte Token hat, liegt der Anteil für *all(s+c)* nur noch bei 22,1 % und für PubMedDS bei gerade einmal 2,2%. Die Länge der Textmarkierungen liegt bei den drei Ansätzen mit jeweils 1,2 Token deutlich unter den 1,6 Token von MedMentions. MedMentions besitzt also im Vergleich zu den anderen Datensätzen eine höhere Dichte an Konzeptmarkierungen und die Markierungen umfassen im Schnitt mehr Token.

Schlussfolgerung: Abschließend lässt sich für den ersten Teil zur Beantwortung der dritten Forschungsfrage - 'Welchen Mehrwert liefert der vorgestellte Korpus im Vergleich zu anderen Datensätzen?' - festhalten, dass die neu vorgestellten Datensätze allein schon durch die Anzahl der Dokumente und markierten UMLS-Konzepte Eigenschaften besitzen, die es bis dato nicht gibt. PubMedDS besitzt zwar ebenfalls eine große Anzahl an Dokumenten, aber mit einem Durchschnittswert von 12,1 UMLS-Konzepten pro Dokument deutlich weniger Konzeptmarkierungen, die genutzt werden können. Diese Problematik kommt vor allem bei Systemen zum Tragen, die eine große Zahl an Markierungen, also einen hohen Recall, benötigen. Auch im Vergleich zur Markierungsdichte von MedMentions bieten die neuen Datensätze eine gute Alternative. Wird von der Qualität der Markierungen abgesehen, liefert vor allem *min1(s+q+c)* ähnliche Werte bei 'Anteil der annotierten Token' und 'annotierte Token / Dokument'.

4.3.2 Aufteilung der semantischen Typen

Für den zweiten Teil der deskriptiven Statistiken wird für dieselben vier Datensätze die Zusammensetzung der semantischen Typen ihrer Konzeptmarkierungen verglichen. Dafür werden alle 127 semantischen Typen des UMLS einer von 24 Übergruppen zugeordnet. Die Gruppen sind aus der Arbeit von Vashishth et al. [36] übernommen. Dort ist auch eine Tabelle zu finden, die für jeden der 127 Typen die zugehörige Übergruppierung zeigt. Abbildung 4.3 zeigt die Zusammensetzung der semantischen Typen in Form von Kreisdiagrammen auf. Um die Datensätze besser vergleichen zu können, werden die Häufigkeiten prozentual zur Gesamtmenge dargestellt.

Auffällig ist, dass sich die Datensätze MedMentions, *all(s+c)* und *min1(s+q+c)* in ihrer Zusammensetzung sehr ähnlich sind. Bis auf kleinere

[4]Die 24 semantischen Gruppen sind übernommen von Vashishth et al. [36].

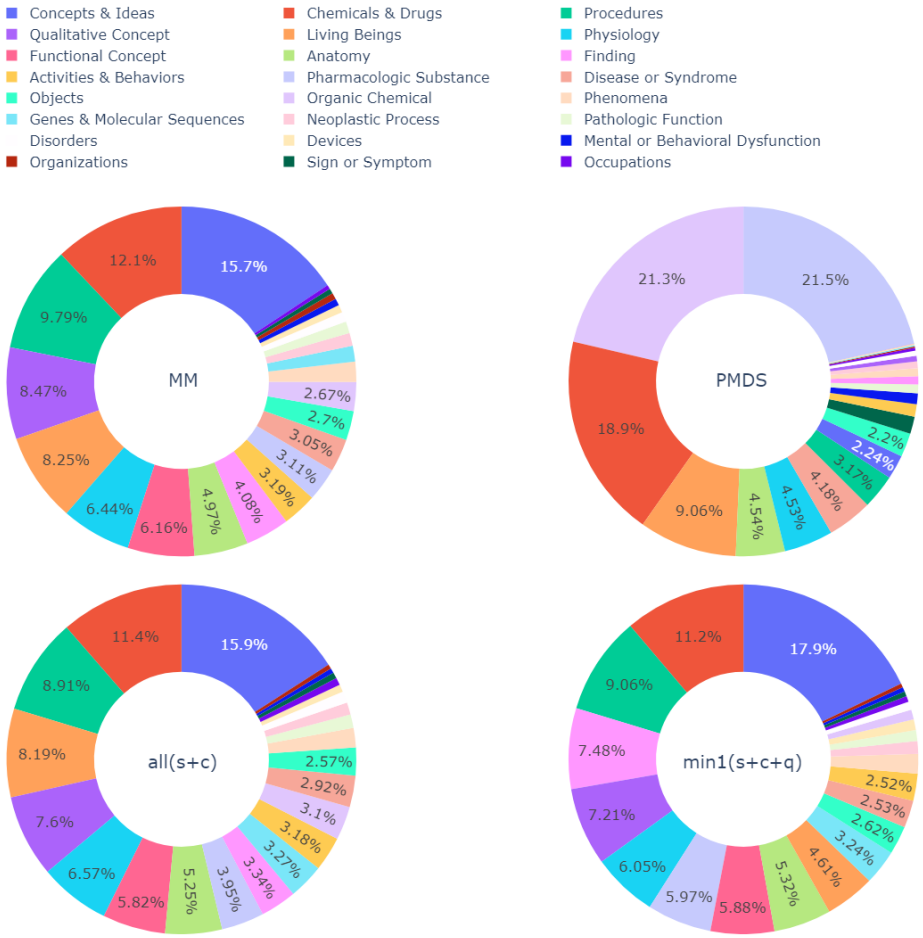

Abbildung 4.3: Darstellung der Verteilung der Konzeptmarkierungen auf die 24 semantischen Gruppen zum Vergleich der zwei ausgewählten Datensätze *min1(s+q+c)* und *all(s+c)* mit MedMentions (MM) und PubMedDS (PMDS).[4]

Abweichungen sind die Rangfolgen der semantischen Gruppen bei diesen drei Datensätzen gleich aufgebaut. PubMedDS auf der anderen Seite unterscheidet sich deutlich von den drei anderen Datensätzen. Die häufigste Gruppe bei den drei sich ähnelnden Datensätzen ist die der 'Concept & Ideas' mit 15,7% bei MedMentions, 15,9% bei *all(s+c)* und 17,9% bei *min1(s+q+c)*. Bei PubMedDS ist diese Gruppe mit gerade einmal 2,24% vertreten und damit kaum im Datensatz vorhanden. Die zweithäufigste Gruppe bei den drei sich ähnelnden Datensätzen ist die der 'Chemicals & Drugs'. Diese macht bei MedMentions 12,1% aus. Bei *all(s+c)* sind es 11,4%. Bei *min1(s+q+c)* sind es 11,2%. Bei PubMedDS sind die 'Chemicals & Drugs' zwar nur die dritthäufigste Gruppe, aber mit 18,9% aller Markierungen ist der Gesamtanteil größer

als bei den anderen Datensätzen. Eine Gruppe, deren Häufigkeit sich in allen vier Datensätzen ähnelt, ist die der 'Living Beings'. Diese Gruppe bewegt sich zwischen 9,06% (PubMedDS) und 4,61% (*min1(s+q+c)*). Der Rest der Übergruppen ist aufgrund der geringen Gesamtmenge eher uninteressant.

Schlussfolgerung: Die Erkenntnisse aus der Analyse liefern den zweiten Teil zur Beantwortung der dritten Forschungsfrage - 'Welchen Mehrwert liefert der vorgestellte Korpus im Vergleich zu anderen Datensätzen?'. Die Erkenntnisse sind, dass die neu erstellten Datensätze in ihrer Zusammensetzung eine breite Streuung an unterschiedlichsten semantischen Typen bieten. Der Vergleichsdatensatz PubMedDS besteht im Gegensatz dazu zu ca. 60% aus den drei Gruppen 'Pharmacologic Substance', 'Organic Chemical' und 'Chemicals & Drugs'. Die genannten Gruppen sind allesamt Substanzen und Chemikalien und ähneln sich daher stark untereinander. Ein Grund für die Einseitigkeit von PubMedDS könnte in der Methodik zur Datensatzerstellung liegen. Die Methodik führt dazu, dass nur Konzepte in den Datensatz aufgenommen werden können, die den MeSH-Termen zugehörig sind. Eine Überprüfung ergibt, dass 83,8 % der in MeSH enthaltenen UMLS-Konzepte einer der drei Gruppen 'Chemicals & Drugs', 'Organic Chemical' oder 'Pharmacologic Substance' zugeordnet werden können. Außerdem gehören 10,7% der Gruppe 'Living Beings' an. Dadurch ist die Band-Breite des PubMedDS-Datensatzes deutlich eingeschränkt. Da die Entity-Linker, die zur Erstellung der neuen Datensätze verwendet wurden, in der Regel MedMentions zum Training nutzen, ist es logisch, dass die neuen Datensätze MedMentions ähneln.

Kapitel 5

Diskussion

Vorgestellt wurde ein Korpus von mehreren mit UMLS-Entitäten markierten Datensätzen. Das folgende Kapitel dient zur Diskussion der Methoden und Ergebnisse. Zunächst wird die Methodik kritisch betrachtet und aufgezeigt, wo deren Schwachpunkte liegen. Dieser Schritt beinhaltet auch das Vorstellen von zukünftigem Arbeitsbedarf. Anschließend werden mögliche Einsatzmöglichkeiten des Korpus diskutiert.

5.1 Limitationen und zukünftige Arbeit

Dieser Abschnitt zeigt Limitationen auf und beschreibt die zukünftige Arbeit, die nötig ist, um den Korpus nützlicher machen zu können.

5.1.1 Qualität der Entity-Linker

Die erste Limitation besteht in der Qualität der Entity-Linker. Die drei Entity-Linker scispaCy, cTAKES und QuickUMLS wurden für diese Arbeit auch aufgrund der leichten Zugänglichkeit und Einfachheit in der Anwendung ausgewählt. Es existieren allerdings auch noch weitere Systeme in der Biomedizin, die genutzt werden könnten und bessere Evaluations-Werte erreichen als die drei ausgewählten. Vor allem mit QuickUMLS war es nicht möglich, wettbewerbsfähige Ergebnisse zu erreichen. Die Stärke des Ensemble-Ansatzes kommt aber vor allem dann zum Tragen, wenn mehrere Entity-Linker von hoher Qualität miteinander verknüpft werden. In Tabelle 4.1 ist zu sehen, dass alle Ansätze (ausgenommen *min2(s+q+c)*), bei denen QuickUMLS verwendet wurde, in den Fβ-scores deutlich schlechter abschneiden als der Rest der Ansätze. Aus dieser Limitation leitet sich ein erster Vorschlag für zukünftige Arbeiten ab. Es wäre denkbar, die Entity-Linker teilweise auszutauschen oder die Gesamtzahl der Entity-Linker zu erhöhen. Denkbare Entity-Linker wie beispielsweise

ArboEL sind bereits in Abschnitt 2.2 beschrieben worden.

5.1.2 Qualität der Evaluation

Ein zweiter Kritikpunkt am Vorgehen in dieser Arbeit betrifft die Evaluation. Zunächst ist zu bemängeln, dass mit MedMentions nur ein Datensatz zur Evaluation genutzt wurde. Das lag daran, dass kein zweiter Datensatz mit markierten UMLS-Konzepten gefunden wurde. Dadurch fehlen Vergleichswerte, um die Evaluationsergebnisse richtig einordnen zu können. Dies wurde versucht mittels einer händischen Zweitevaluation auszugleichen. Aber auch diese wurde nicht wissenschaftlich korrekt durchgeführt. Aufgrund von mangelnden Ressourcen musste die Evaluation in Eigenarbeit realisiert werden. Daher konnte nur eine einzige Methode und auch nur ein kleiner Ausschnitt von MedMentions evaluiert werden. Besser wäre es, ein Team unabhängiger UMLS-Experten zusammenzustellen, welche dann die Evaluation umfangreicher und ohne Bias durchführen könnten.

5.1.3 Erweiterung durch Relationen

Um den Datensatz aufzuwerten und für eine größere Menge an Modellen nutzbar zu machen, können die Beziehungen der UMLS-Entitäten eingebaut werden. Wie in Unterabschnitt 3.1.1 beschrieben, handelt es sich beim UMLS um einen Wissensgraph. Dieser Wissensgraph beinhaltet zusätzlich zu den Konzepten auch die Beziehungen der Konzepte zueinander. Diese Beziehungen können in den Datensatz integriert werden. Die größte Aufgabe besteht dabei darin, mögliche Entitätenpaare für eine Beziehung zu identifizieren. Die einfachste und wohl sinnvollste Möglichkeit wäre, in jedem Dokument des Datensatzes alle markierten Entitäten zu ermitteln. Anschließend wird jede einzelne dieser Entitäten mit allen anderen Entitäten des Dokuments gepaart und im UMLS auf eine Beziehung überprüft. Dadurch werden alle Beziehungen ermittelt, die in diesem Dokument vorhanden sind. Das hat den Vorteil, dass bei der Entwicklung einer neuen Methode, die diesen Datensatz nutzt, selbst entschieden werden kann, ob die Beziehungen noch eingeschränkt werden sollen. Zum Beispiel könnte eine maximale Entfernung der beiden Konzepte zueinander festgelegt werden.

5.2 Einsatzmöglichkeiten

Nachfolgend werden vier Ideen aufgezählt, bei denen der Korpus möglicherweise eingesetzt werden kann:

1. Eine Einsatzmöglichkeit des Korpus ist als Datengrundlage in einer Suchmaschine. Die markierten UMLS-Konzepte könnten – ähnlich den MeSH-Termen in PubMed – als Indexterme dienen. Der Nutzen läge dann im Filtern von Dokumenten.

2. Als zweite Einsatzmöglichkeit ist das Training eines Entity-Linking-Modells denkbar. Durch die vorgestellten Ensemble-Methoden werden die Ergebnisse mehrerer Entity-Linker zusammengefasst, wodurch die Einzelergebnisse verbessert werden können. Wird nun auf den daraus entstandenen Datensätzen ein neuer Entity-Linker trainiert, kann möglicherweise eine Verbesserung der Linking-Resultate erreicht werden.

3. Als weitere Idee kann ein Sprachmodell entwickelt werden, das durch das zusätzliche Wissen des UMLS-Graphen angereichert wird. Colon-Hernandez et al. [8] stellen einige Möglichkeiten vor. Es wird zwischen Input-, Architecture- und Output-Injection unterschieden. Input-Injection liegt bei Modellen vor, die das Wissen des Graphen bereits in den Eingabedaten integrieren (z.B. COMmonsEnse Transformers (COMET) [7]). Bei Architecture-Injection wird die Architektur durch spezielle Schichten oder Module angepasst, um den Wissensgraph nutzen zu können (z.B. KnowBERT [28]). Unter die Output-Injection fallen Modelle, die den Wissensgraphen nutzen, um die Ausgabe anzupassen (z.B. SemBERT [38]). Jede dieser Möglichkeiten ist auch für die biomedizinische Domäne mit den UMLS als Wissensgraph denkbar. Dafür wäre allerdings die Integration der UMLS-Relationen in die Datensätze nötig.

4. Es ist vorstellbar, aus einem Teil des Korpus einen neuen Datensatz zur PICO-Extraktion zu erstellen. PICO steht für Patient, Intervention, Comparison und Outcome. Ein Datensatz zur PICO-Extraktion müsste in Randomized Controlled Trials (RCTs) die Textstellen markiert haben, die einem der PICO-Elemente angehören. Die Idee ist es, aus dem Korpus die Artikel herauszusuchen, die RCTs sind und gleichzeitig in einer Cochrane Review[1] verwendet wurden. In Cochrane Reviews gibt es oft Tabellen, in denen die PICO-Elemente der darin verwendeten RCTs aufgezeigt werden. Es ist denkbar, in diesen Tabellen ebenfalls UMLS-Konzepte zu extrahieren und dann die Konzepte der Tabellen mit den Konzepten der RCTs zu vergleichen und dadurch in den RCT-Texten PICO-Elemente zu markieren.

[1] Mehr Informationen zu Cochrane und RCTs sind unter https://www.cochranelibrary.com/about/about-cochrane-reviews zu finden. Abgerufen am 12.06.2023

Kapitel 6

Zusammenfassung

Die vorgestellte Arbeit zeigt ein Konzept auf, das zur Erstellung großer Datensätze mit extrahierten biomedizinischen Konzepten dient. Die entwickelte Methodik sieht vor, mittels Ensemble-Methoden die Ergebnisse mehrerer Entity-Linker zusammenzuführen und somit deren einzelne Stärken zu nutzen und gleichzeitig die Schwächen zu verringern. Genutzt wurden ca. 21 Mio. Abstracts von biomedizinischen Veröffentlichungen. In den Texten wurden zunächst mittels der Entity-Linker scispaCy, cTAKES und QuickUMLS UMLS-Konzepte markiert. Anschließend wurden durch die drei Ensemble-Methoden *min1*, *min2* und *all* die Ergebnisse zusammengeführt. Durch die Kombination aus Entity-Linkern und Ensemble-Methoden wurden insgesamt neun Datensätze erstellt. Zu Beginn der Arbeit wurden drei Forschungsfragen formuliert. Diese wurden später durch verschiedene Evaluationsschritte beantwortet.

Die **erste Forschungsfrage** lautete 'Welche Kombinationen aus Ensemble-Methode und Entity-Linker-Auswahl führen zu den besten Ergebnissen?'. Zur Beantwortung wurden für die verschiedenen Datensätze die Kennwerte Precision, Recall, F0.5-, F1- und F2-score gegenübergestellt und verglichen. In der Regel bieten die Fβ-Werte eine gute Bewertungsgrundlage, da Precision und Recall gegeneinander aufgewogen werden. F0.5 gewichtet den Recall halb so stark wie die Precision, F1 gewichtet beide gleich und F2 gewichtet den Recall doppelt so stark. Bei der Bewertung der Ergebnisse kommt es dann auf die persönliche Relevanz an. Der beste Ansatz bei F0.5 und F1 war *all(s+c)* mit 0,470 und 0,430. Dieser ist vor allem dann zu bevorzugen, wenn ein Datensatz gesucht wird, der ein gutes Gleichgewicht zwischen Precision und Recall bietet. Bei F2 war der beste Ansatz *min1(s+c)* mit 0,440. Dieser ist eine gute Option für Methoden, die einen Datensatz mit etwas Fokus auf den Recall benötigen. Auffällig bei der Evaluation war außerdem, dass QuickUMLS im Vergleich zu scispaCy und cTAKES deutlich schlechter abgeschnitten hat. Das führt da-

zu, dass auch die meisten Ensembles, bei denen QuickUMLS genutzt wurde, schlechte Werte liefern.

Die **zweite Forschungsfrage** lautete 'Sind die Precision-Werte der Evaluation aufgrund von ungleichen Bedingungen bei der Erstellung der Datensätze verfälscht?'. Um diese Frage zu beantworten, wurde eine händische Zweitevaluation vorgenommen. Da die Zweitevaluation von Hand und ohne externe Hilfe durchgeführt wurde, bestand die Gefahr, einen persönlichen Bias einfließen zu lassen. Um dem entgegenzuwirken, wurden strikte Regeln aufgestellt, die bei der Bewertung der einzelnen Konzepte angewendet wurden. Aufgrund der begrenzten Ressourcen beschränkte sich die Evaluation auf das *all(s+c)* Verfahren. Für 100 zufällig ausgewählte Abstracts wurde von Hand jede in *all(s+c)* markierte Textstelle bewertet und neu in TP oder FP eingeordnet. Anschließend wurde die Precision neu berechnet. Die Ergebnisse der Zweitevaluation legen nahe, dass die Precision um ca. 14% höher sein könnte, als es die Ergebnisse der ursprünglichen Evaluation andeuten.

Die **dritte Forschungsfrage** lautete 'Welchen Mehrwert liefert der vorgestellte Korpus im Vergleich zu anderen Datensätzen?'. Der Mehrwert wurde durch verschiedene deskriptive Statistiken aufgezeigt. Dabei wurde ein Vergleich zu den Datensätzen MedMentions und PubMedDS gezogen. Zu Beginn wurden allgemeine Kennwerte verglichen. Die wichtigste Erkenntnis daraus besteht im Umfang der neuen Datensätze. Diese sind mit 21.168.134 Dokumenten ca. 4.800-mal größer als MedMentions mit 4.392 Dokumenten und fast doppelt so groß wie PubMedDS mit 13.197.430 Dokumenten. Außerdem sind bei *min1(s+q+c)* mit bis 803.483 einzigartigen UMLS-Konzepten 23-mal mehr Konzepte als in MedMentions (34.724 einzigartige Konzepte) und 18-mal so viele wie in PubMedDS (44.881 einzigartige Konzepte) vorhanden. Im Anschluss an die allgemeinen Kennwerte wurden die markierten Konzepte und deren semantische Typen genauer betrachtet. Die Erkenntnis aus dieser Analyse war, dass die neu erstellten Datensätze eine breite Streuung in den semantischen Typen der Konzepte besitzen, wobei der größte Anteil den Gruppen 'Concepts & Ideas' (15,9% bei *all(s+c)*), 'Chemical & Drugs' (11,4% bei *all(s+c)*) und 'Procedures' (8,91% bei *all(s+c)*) zuzusprechen ist. Die Zusammensetzung ähnelt stark der Zusammensetzung bei MedMentions. PubMedDS hingegen ist aufgrund der Methodik auf Konzepte des MeSH-Vokabulars beschränkt und besteht zu mehr als 60% aus Konzepten der Gruppen 'Pharmacologic Substance', 'Organic Chemical' und 'Chemicals & Drugs'. Der Mehrwert der vorgestellten Arbeit besteht also darin, Datensätze erstellt zu haben, die in ihrer Konzeptzusammensetzung eine ähnliche Struktur wie MedMentions besitzen und gleichzeitig größer als PubMedDS sind.

Zum Schluss wurde im Diskussions-Abschnitt der Nutzen der Arbeit aufgezeigt. Zuerst wurden mögliche Limitationen genannt und zukünftige Arbeiten diskutiert. Ein limitierender Faktor ist die Qualität der Entity-Linker und dabei insbesondere von QuickUMLS. Daher wurde vorgeschlagen, weitere Entity-Linker auszutesten, um die Qualität der Datensätze zu verbessern. Ein weiterer Kritikpunkt an dieser Arbeit lag im Vorgehen bei der Evaluation. Erstens wurde mit MedMentions nur ein Datensatz zur Evaluation genutzt. Dadurch fehlen Vergleichswerte, um die Evaluationsergebnisse richtig bewerten zu können. Zweitens wurde die händische Zweitevaluation in Eigenarbeit durchgeführt, weshalb ein persönlicher Bias nicht auszuschließen ist. Um das zu beheben, müsste in Zukunft eine Gruppe von UMLS-Experten zusammengestellt werden, welche die händische Zweitevaluation unabhängig durchführen. Als letzter Zukunftsausblick wurde vorgeschlagen, im Korpus die Beziehungen der UMLS-Konzepte mit einzubauen. Dies würde die Anzahl der Nutzungsmöglichkeiten stark erweitern. Anschließend wurden vier Einsatzmöglichkeiten für den Korpus vorgeschlagen. Die Datensätze könnten für die Entwicklung einer Suchmaschine oder zum Training von neuen Entity-Linkern verwendet werden. Eine weitere Möglichkeit ist das Entwickeln von Sprachmodellen, die in ihrem Training durch Wissensgraphen angereichert werden. Als letzte Idee könnte der vorgestellte Korpus dazu dienen, einen Datensatz zur PICO-Extraktion zu erstellen.

Literaturverzeichnis

[1] Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. Entity linking via explicit mention-mention coreference modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4644–4658. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.343. URL `https://doi.org/10.18653/v1/2022.naacl-main.343`.

[2] Liz Amos, David Anderson, Stacy Brody, Anna Ripple, and Betsy L. Humphreys. UMLS users and uses: a current overview. *J. Am. Medical Informatics Assoc.*, 27(10):1606–1611, 2020. doi: 10.1093/jamia/ocaa084. URL `https://doi.org/10.1093/jamia/ocaa084`.

[3] Alan R. Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *J. Am. Medical Informatics Assoc.*, 17(3):229–236, 2010. doi: 10.1136/jamia.2009.002733. URL `https://doi.org/10.1136/jamia.2009.002733`.

[4] Kimberly Van Auken, Mary L. Schaeffer, Peter McQuilton, Stanley J. F. Laulederkind, Donghui Li, Shur-Jen Wang, G. Thomas Hayman, Susan Tweedie, Cecilia N. Arighi, James Done, Hans-Michael Müller, Paul W. Sternberg, Yuqing Mao, Chih-Hsuan Wei, and Zhiyong Lu. BC4GO: a full-text corpus for the biocreative IV GO task. *Database J. Biol. Databases Curation*, 2014, 2014. doi: 10.1093/database/bau074. URL `https://doi.org/10.1093/database/bau074`.

[5] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3613–3618. Asso-

ciation for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1371. URL https://doi.org/10.18653/v1/D19-1371.

[6] Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue): 267–270, 2004. doi: 10.1093/nar/gkh061. URL https://doi.org/10.1093/nar/gkh061.

[7] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4762–4779. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1470. URL https://doi.org/10.18653/v1/p19-1470.

[8] Pedro Colon-Hernandez, Catherine Havasi, Jason B. Alonso, Matthew Huggins, and Cynthia Breazeal. Combining pre-trained language models and structured knowledge. *CoRR*, abs/2101.12294, 2021. URL https://arxiv.org/abs/2101.12294.

[9] The Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic Acids Res.*, 47(Database-Issue):D330–D338, 2019. doi: 10.1093/nar/gky1055. URL https://doi.org/10.1093/nar/gky1055.

[10] A. C. Davison, D. V. Hinkley, and G. A. Young. Recent developments in bootstrap methodology. *Statistical Science*, 18(2):141–157, 2003. doi: 10.1214/ss/1063994969. URL https://www.jstor.org/stable/3182844.

[11] Dina Demner-Fushman, Willie J. Rogers, and Alan R. Aronson. Metamap lite: an evaluation of a new java implementation of metamap. *J. Am. Medical Informatics Assoc.*, 24(4):841–844, 2017. doi: 10.1093/jamia/ocw177. URL https://doi.org/10.1093/jamia/ocw177.

[12] Rezarta Islamaj Dogan and Zhiyong Lu. An improved corpus of disease mentions in pubmed citations. In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, BioNLP@HLT-NAACL Montrèal, Canada, June 8, 2012*, pages 91–99. Association for Computational Linguistics, 2012. URL https://aclanthology.org/W12-2411/.

[13] Armen Yuri Gasparyan, Marlen Yessirkepov, Alexander Voronov, Anna Koroleva, and George Kitas. Comprehensive approach to open access publishing: Platforms and tools. *Journal of Korean Medical Science*, 34

(27):e184, 2019. doi: 10.3346/jkms.2019.34.e184. URL `https://doi.org/10.3346/jkms.2019.34.e184`.

[14] M. Gersenovic. The icd family of classifications. *Methods of information in medicine*, 34(1-2):172–175, 1995. URL `http://www.ncbi.nlm.nih.gov/pubmed/908212`.

[15] Cyril Goutte and Éric Gaussier. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *Advances in Information Retrieval, 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005, Proceedings*, volume 3408 of *Lecture Notes in Computer Science*, pages 345–359. Springer, 2005. doi: 10.1007/978-3-540-31865-1_25. URL `https://doi.org/10.1007/978-3-540-31865-1_25`.

[16] Trisha Greenhalgh. How to read a paper: The medline database. *BMJ*, 315(7101):180–183, 1997. doi: 10.1136/bmj.315.7101.180. URL `https://www.bmj.com/content/315/7101/180`.

[17] Nicolas Heist, Sven Hertling, Daniel Ringler, and Heiko Paulheim. Knowledge graphs on the web - an overview. In *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*, volume 47 of *Studies on the Semantic Web*, pages 3–22. IOS Press, 2020. doi: 10.3233/SSW200009. URL `https://doi.org/10.3233/SSW200009`.

[18] Martin Komenda, Daniel Schwarz, Jan Svancara, Christos Vaitsis, Nabil Zary, and Ladislav Dusek. Practical use of medical terminology in curriculum mapping. *Comput. Biol. Medicine*, 63:74–82, 2015. doi: 10.1016/j.compbiomed.2015.05.006. URL `https://doi.org/10.1016/j.compbiomed.2015.05.006`.

[19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.703. URL `https://doi.org/10.18653/v1/2020.acl-main.703`.

[20] Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. Biocreative V CDR task corpus:

a resource for chemical disease relation extraction. *Database J. Biol. Databases Curation*, 2016, 2016. doi: 10.1093/database/baw068. URL https://doi.org/10.1093/database/baw068.

[21] D. A. B. Lindberg, B. L. Humphreys, and A. T. McCray. The unified medical language system. *Yearbook of medical informatics*, 2(1):41–51, 1993. doi: 10.1055/s-0038-1637976. URL https://doi.org/10.1055/s-0038-1637976.

[22] Daniel Loureiro and Alípio Mário Jorge. Medlinker: Medical entity linking with neural representations and dictionary matching. In *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II*, volume 12036 of *Lecture Notes in Computer Science*, pages 230–237. Springer, 2020. doi: 10.1007/978-3-030-45442-5_29. URL https://doi.org/10.1007/978-3-030-45442-5_29.

[23] David T. Marc and Saif S. Khairat. Medical subject headings (mesh) for indexing and retrieving open-source healthcare data. In *Integrating Information Technology and Management for Quality of Care [ICIMTH 2014, Athens, Greece, 10-13 July 2014]*, volume 202 of *Studies in Health Technology and Informatics*, pages 157–160. IOS Press, 2014. doi: 10.3233/978-1-61499-423-7-157. URL https://doi.org/10.3233/978-1-61499-423-7-157.

[24] Johanna McEntyre and David Lipman. Pubmed: bridging the information gap. *CMAJ: Canadian Medical Association Journal*, 164(9):1317–1319, 2001. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC81025.

[25] Sunil Mohan and Donghui Li. Medmentions: A large biomedical corpus annotated with UMLS concepts. In *1st Conference on Automated Knowledge Base Construction, AKBC 2019, Amherst, MA, USA, May 20-22, 2019*, 2019. doi: 10.24432/C5G59C. URL https://doi.org/10.24432/C5G59C.

[26] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. Scispacy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 319–327. Association for Computational Linguistics, 2019. doi: 10.18653/v1/w19-5034. URL https://doi.org/10.18653/v1/w19-5034.

[27] Naoaki Okazaki and Jun'ichi Tsujii. Simple and efficient algorithm for approximate dictionary matching. In *COLING 2010, 23rd International*

Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2010, Beijing, China, pages 851–859. Tsinghua University Press, 2010. URL `https://aclanthology.org/C10-1096/`.

[28] Matthew E. Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 43–54. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1005. URL `https://doi.org/10.18653/v1/D19-1005`.

[29] Sameer Pradhan, Noémie Elhadad, Brett R. South, David Martínez, Lee M. Christensen, Amy Vogel, Hanna Suominen, Wendy W. Chapman, and Guergana Savova. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *J. Am. Medical Informatics Assoc.*, 22(1):143–154, 2015. doi: 10.1136/amiajnl-2013-002544. URL `https://doi.org/10.1136/amiajnl-2013-002544`.

[30] Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin Kipper Schuler, and Christopher G. Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *J. Am. Medical Informatics Assoc.*, 17(5):507–513, 2010. doi: 10.1136/jamia.2009.001560. URL `https://doi.org/10.1136/jamia.2009.001560`.

[31] Eric W. Sayers, Evan E. Bolton, J. Rodney Brister, Kathi Canese, Jessica Chan, Donald C. Comeau, Ryan Connor, Kathryn Funk, Chris Kelly, Sunghwan Kim, Tom Madej, Aron Marchler-Bauer, Christopher J. Lanczycki, Stacy Lathrop, Zhiyong Lu, Françoise Thibaud-Nissen, Terence D. Murphy, Lon Phan, Yuri Skripchenko, Tony Tse, Jiyao Wang, Rebecca Williams, Barton W. Trawick, Kim D. Pruitt, and Stephen T. Sherry. Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, 50(D1):20–26, 2022. doi: 10.1093/nar/gkab1112. URL `https://doi.org/10.1093/nar/gkab1112`.

[32] Conrad L. Schoch, Stacy Ciufo, Michael Domrachev, Carol L. Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard McVeigh, Kathleen O'Neill, Barbara Robbertse, Shobha Sharma, Vladimir Soussov, John P. Sullivan, Lu Sun, Seán Turner, and Ilene Karsch-Mizrachi. NCBI taxonomy: a comprehensive update on curation, resources and tools. *Da-*

tabase J. Biol. Databases Curation, 2020, 2020. doi: 10.1093/database/baaa062. URL `https://doi.org/10.1093/database/baaa062`.

[33] Ariel S. Schwartz and Marti A. Hearst. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proceedings of the 8th Pacific Symposium on Biocomputing, PSB 2003, Lihue, Hawaii, USA, January 3-7, 2003*, pages 451–462, 2003. URL `http://psb.stanford.edu/psb-online/proceedings/psb03/schwartz.pdf`.

[34] Juliet Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46(1):561–584, 1995. doi: 10.1146/annurev.ps.46.020195.003021. URL `https://doi.org/10.1146/annurev.ps.46.020195.003021`.

[35] Luca Soldaini and Nazli Goharian. Quickumls: a fast, unsupervised approach for medical concept extraction. *MedIR Workshop, SIGIR 2016*, 2016.

[36] Shikhar Vashishth, Denis Newman-Griffis, Rishabh Joshi, Ritam Dutt, and Carolyn P. Rosé. Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. *J. Biomed. Informatics*, 121:103880, 2021. doi: 10.1016/j.jbi.2021.103880. URL `https://doi.org/10.1016/j.jbi.2021.103880`.

[37] Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. Biobart: Pretraining and evaluation of A biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing, BioNLP@ACL 2022, Dublin, Ireland, May 26, 2022*, pages 97–109. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.bionlp-1.9. URL `https://doi.org/10.18653/v1/2022.bionlp-1.9`.

[38] Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. Semantics-aware BERT for language understanding. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9628–9635. AAAI Press, 2020. URL `https://ojs.aaai.org/index.php/AAAI/article/view/6510`.

[39] Jinguang Zheng, Daniel Howsmon, Boliang Zhang, Juergen Hahn, Deborah L. McGuinness, James A. Hendler, and Heng Ji. Entity linking for biomedical literature. *BMC Medical Informatics Decis. Mak.*, 15-S(S-1):

S4, 2015. doi: 10.1186/1472-6947-15-S1-S4. URL https://doi.org/10.1186/1472-6947-15-S1-S4.

Erklärung

Ich versichere, dass ich die vorliegende Arbeit selbstständig und nur unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt habe, insbesondere sind wörtliche oder sinngemäße Zitate als solche gekennzeichnet. Mir ist bekannt, dass Zuwiderhandlung auch nachträglich zur Aberkennung des Abschlusses führen kann. Ich versichere, dass das elektronische Exemplar mit den gedruckten Exemplaren übereinstimmt.

Untergassen, 17. Juni 2023

. . . . . .

Gregor Pfänder