Annotation und Management heterogener medizinischer Studienformulare

Victor Christen
Institut für Informatik, Universität Leipzig
christen@informatik.uni-leipzig.de

ABSTRACT

Medizinische Formulare werden für die Dokumentation innerhalb der klinischen Forschung oder der Dokumentation von Patientendaten verwendet. Es existiert eine Vielzahl verschiedener Formulare, die für verschiedene Nutzungszwecke bzw. Anwendungen erstellt werden. Aufgrund der resultierenden Heterogenität ist eine Vergleichbarkeit, eine studienübergreifende Analyse oder eine effiziente Suche nicht ohne weiteres möglich. Um die Interoperabilität der Anwendungen, die auf der Auswertung von Formularen basieren, zu erhöhen, ist eine einheitliche Annotation von medizinischen Formularen mittels einer medizinischen Wissensbasis hilfreich. Eine solche Wissensbasis ist das Unified Medical Language System(UMLS), welches biomedizinisch relevante Konzepte umfasst. Diese Arbeit befasst sich mit der semi-automatischen Annotation von Studienformularen. Basierend auf einem allgemeinen Matching-Workflow, werden weitere Lösungsansätze präsentiert, um die Besonderheiten der Annotation von Studienformularen zu behandeln.

Keywords: semantische Annotationen, medizinische Formulare, klinische Studien, UMLS

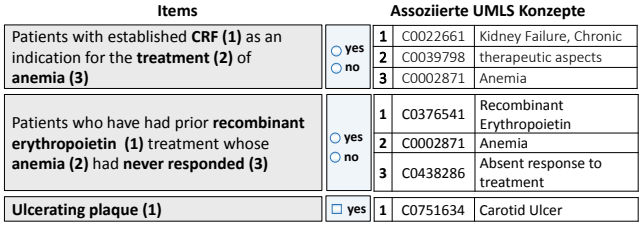
1. EINLEITUNG

Medizinische Formulare werden verwendet, um Patientendaten und resultierende Daten innerhalb einer klinischen Studie zu dokumentieren. So werden Studienformulare für die Rekrutierung der Probanden der jeweiligen Studien verwendet, indem die Ein- und Ausschlusskriterien definiert werden. Momentan sind $\sim$ 180000 Studien auf `http://clinicaltrials.gov` registriert, wobei jede Studie eine Menge von Case Report Forms (CRF) umfasst, um die notwendigen Daten zu dokumentieren. Im Allgemeinen werden Formulare einer Studie neu erstellt ohne bereits existierende Formulare wieder zu verwenden.

Aufgrund der hohen Anzahl heterogener Formulare ist eine studienübergreifende Analyse oder der Datenaustausch komplex und nicht ohne weiteres effizient realisierbar. Um eine einheitliche und strukturierte Repräsentation zu ermöglichen werden die Formulare mit Konzepten von standardisierten Vokabularen wie z.B. Ontologien annotiert [4]. Ontologien sind in der Biomedizin für die Anreicherung von Realweltobjekten weit verbreitet. Die Gene Ontology (GO) wird verwendet, um die Funktionen von Genen und Proteinen zu beschreiben, mithilfe der Medical Subject Headings (MeSH) [8] Ontologie werden wissenschaftliche Publikationen annotiert, und durch die Annotation mit Konzepten der SNOMED CT Ontologie [3] ist eine strukturierte und einheitliche Verwaltung von Patientendaten möglich. Das UMLS [1] repräsentiert eine biomedizinische Wissensbasis, die mehr als 100 biomedizinische Ontologien integriert, wie z.B. SNOMED CT, National Cancer Institute Thesaurus(NCIT) oder MeSH und umfasst $\sim$ 2.8 Millionen Konzepte. Die verschiedenen Anwendungsfälle zeigen das Potential für die Vereinfachung der semantischen Suche und der Datenintegration durch die Annotation von Realweltobjekten mittels der Konzepte von Ontologien. Die Annotation von Formularen hat folgenden Mehrwert:

- **Studienübergreifende Analysen** Eine studienübergreifende Analyse umfasst Studien mit einer ähnlichen Thematik. Die Identifikation ähnlicher Studien ist mithilfe der annotierten Formulare bzgl. der Studien effizient und effektiv durchführbar. Ein Beispiel für eine studienübergreifende Analyse ist der Vergleich der Wirksamkeit und Sicherheit von medikamentbeschichteten Stents und unbeschichteten Stents für Herzkranzgefäße [7]. Bei dieser Analyse wurden 9470 Patienten von 22 randomisierten kontrollierten Studien und 182901 Patienten von 34 Beobachtungsstudien betrachtet. Bei der Auswertung der Daten müssen die Antworten der Fragen der Formulare integriert werden. Die Annotationen der Formulare können für den Integrationsprozess verwendet werden, indem initial durch die Annotationen ähnliche Items identifiziert werden. Die Daten, die die ähnlichen Items betreffen, werden durch weitere Integrationsschritte vereinheitlicht, so dass eine Analyse möglich ist.
- **Erstellung von Formularen** Bisher werden Formulare mit ihren Items für eine durchzuführende Studie neu erstellt. Die Erstellung eines Formulars ist ein aufwändiger Prozess, da z.B. eine unscharfe Formulierung der Ein- und Ausschlusskriterien zu einer mögl. Menge an Probanden führt, die für die Studie nicht vorgesehen waren. Durch die Identifikation bereits annotierte Formulare, die der Thematik der durchzuführenden

27^{th} GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), 26.05.2015 - 29.05.2015, Magdeburg, Germany.



Items		Assoziierte UMLS Konzepte		
Patients with established **CRF (1)** as an indication for the **treatment (2)** of **anemia (3)**	○ yes ○ no	1	C0022661	Kidney Failure, Chronic
		2	C0039798	therapeutic aspects
		3	C0002871	Anemia
Patients who have had prior **recombinant erythropoietin (1)** treatment whose **anemia (2)** had **never responded (3)**	○ yes ○ no	1	C0376541	Recombinant Erythropoietin
		2	C0002871	Anemia
		3	C0438286	Absent response to treatment
Ulcerating plaque (1)	□ yes	1	C0751634	Carotid Ulcer

Figure 1: Beispiel für die Annotation der Items eines Formulars mit Konzepten des UMLS

Studie entsprechen oder ähneln, können ähnliche Items bei der Erstellung des neuen Formulars wiederverwendet werden.

Ein Formular besteht aus einer Menge von *Items*. Ein *Item* umfasst eine Frage und die dazugehörigen Antwortmöglichkeiten. Eine Antwort hat einen Datentyp wie z.B. Boolean oder String, bei Freitextantworten, oder kann durch einen vordefinierten Bereich wie z.B. das Alter von 0 bis 140 oder eine vorgegebene Menge, die z.B. die möglichen Symptome definiert, eingeschränkt werden. Bei der Annotation eines medizinischen Formulars wird jedem Item eine Menge von Konzepten des UMLS zugeordnet, so dass diese semantisch beschrieben sind. Ein Beispiel für die Annotation eines Formulars für die Ein- und Ausschlusskriterien einer Studie bzgl. Blutarmut ist in Abb. 1 dargestellt. Das Beispiel verdeutlicht die Komplexität der automatischen Identifikation von Annotationen, da z.B. wie in Frage 1 signifikante Wortgruppen zu einem Konzept korrespondieren oder die Frage 3 ein Synonym enthält bzgl. des korrespondierenden Konzepts.

Die *Medical Data Models* Plattform bietet bereits Möglichkeiten für die Erstellung, die Analyse, den Austausch und die Wiederverwendung von Formularen in einem zentralen Metadaten Repository [2]. Aktuell umfasst das Repository mehr als 9000 Versionen von medizinischen Formularen und über 300000 Items. Um die semantische Heterogenität zu reduzieren, werden die Formulare mit Konzepten des UMLS annotiert. Die Annotation der Formulare ist im MDM bisher nur manuell durchführbar und somit sind viele Formulare nicht bzw. unvollständig annotiert, da dieser Prozess sehr zeitintensiv ist.

Die automatische Annotation von Formularen ist thematisch verwandt mit dem Ontologie-Matching, das eine Menge von Korrespondenzen, Mapping genannt, zwischen den Konzepten von zwei oder mehreren Ontologien generiert. Dabei repräsentiert eine Korrespondenz eine semantische Ähnlichkeit zwischen zwei Konzepten. Bei der Annotation von Formularen werden ebenfalls Korrespondenzen ermittelt, wobei eine Korrespondenz zwischen einem Item und einem Konzept ist, welches das Item semantisch beschreibt. Auf dem Gebiet des Ontologie-Matchings existieren eine Vielzahl von Verfahren [11], die eine effiziente und effektive Generierung eines Ontologie-Mappings realisieren, wie z.B. GOMMA [6]. Aufgrund dessen werden Ansätze des Ontologie-Matchings für die Annotation von Studienformularen verwendet, wie z.B. diverse String-Matchverfahren oder Blocking-Techniken. Jedoch unterscheiden sich Formulare und Ontologien dahingehend, dass Formulare nicht formal strukturiert sind und aufgrund der besseren Verständlichkeit einen höheren Freitextanteil beinhalten. Die bisherigen Ontologie-Matching Verfahren unterstützen nur unzureichend das Matching von Entitäten mit einem hohen Freitextanteil sowie die Erkennung von n:m Korrespondenzen.

Das Ziel unserer Forschung ist die Verbesserung der Qualität der Annotationen. Des Weiteren soll ein Formular Management System (FMS) realisiert werden, das die Verwaltung der Formulare, Ontologien und der berechneten Annotationen ermöglicht. Das FMS soll zusätzlich das Annotationsverfahren beinhalten sowie Funktionalitäten für die Suche, Analyse und Verifikation der Annotationen von Formularen bereitstellen. Für die Verbesserung der Qualität der Annotationen und der Effizienz der Verfahren sollen folgende Aspekte betrachtet werden.

- **Identifikation von signifikanten Termen und zusammengehörigen Einheiten** Die Fragen innerhalb eines Formulars sind in natürlicher Sprache formuliert. Jedoch sind die Konzepte von Ontologien in einer kompakten Form beschrieben und auf die relevanten Terme beschränkt. Somit ist es notwendig innerhalb einer Frage die signifikanten Terme zu identifizieren. Des Weiteren kann eine Frage aus mehreren semantischen Einheiten bestehen, die jeweils durch ein Konzept beschrieben werden. Aufgrund dessen ist es notwendig diese Wortgruppen zu identifizieren.
- **Wiederverwendung von annotierten Formularen** Da das UMLS eine hohe Anzahl von Konzepten umfasst, ist die vollständige Berechnung des kartesischen Produkts bzgl. aller Fragen eines Formulars sehr zeitintensiv. Durch die Verwendung bereits annotierter Items ist es möglich, die zeitliche Komplexität zu reduzieren, indem zu dem unannotierten Item ähnliche, bereits annotierte Items ermittelt werden. Die assoziierten Konzepte der annotierten Items sind Kandidaten für die Annotation des unannotierten Items.
- **Erweiterte Selektionsstrategien** Beim Ontologie-Matching wird ein Mapping generiert, wobei durch Top-k Selektionsstrategien die Korrespondenzen basierend auf einer berechneten Ähnlichkeit selektiert werden. Da eine Frage durch mehrere Konzepte beschrieben werden kann, die Konzepte jedoch nicht ähnlich sind, sind solche Selektionsstrategien nicht effektiv. Aufgrund dessen sind komplexere Selektionsstrategien erforderlich, die n:m Korrespondenzen berücksichtigen.
- **Verifikationsverfahren** Mithilfe eines Expertenkonsortiums soll die Qualität der Annotationen innerhalb des FMS durch die unterstützte manuelle Verifikation der ermittelten Annotationen erhöht werden. Des Weiteren ist es möglich, dass ein Experte weitere Annotationen vorschlagen kann. Zusätzlich soll ein Verifizierungsverfahren realisiert werden, welches die Widerspruchsfreiheit und die Minimalität der assoziierten Konzepte mit berücksichtigt. So ist z.B. eine Menge von Annotationen nicht korrekt, wenn zwei Konzepte innerhalb dieser Menge als disjunkt definiert sind, dass heißt diese zwei Konzepte besitzen keine gemeinsamen

Instanz. Ein Annotations-Mapping ist nicht minimal, wenn zwei Konzepte dieselbe Thematik beschreiben. Mithilfe der *is_a*-Hierarchie und den Disjunktheitsbeziehungen innerhalb einer Ontologie sind solche Konflikte identifizierbar und durch die Anwendung von Auflösungsstrategien zu beheben.

- **Reduktion der Vergleiche im Annotationsprozess** Aufgrund der hohen Anzahl der Konzepte bei Ontologien ist es sinnvoll die Anzahl der Vergleiche im Annotationsprozess einzuschränken, um eine hohe Effizienz zu erzielen. Es existieren bereits Verfahren, die eine Reduktion der Vergleiche ermöglichen wie z.B. Längenfilter, PPJoin[12] oder Locality Sensitive Hashing (LSH) [5]. Als Ziel unserer Forschung sollen ähnliche Verfahren in den Annotationsprozess integriert werden bzw. neue Verfahren realisiert werden.

Es wurde begonnen ein automatisches Verfahren für die Annotation von Formularen zu implementieren. Die Realisierung eines basalen Workflows und erste Erweiterungen wurden in einer eingereichten Publikation „Annotating Medical Forms using UMLS“ beschrieben. Die Ergebnisse verdeutlichen die Schwierigkeiten der automatischen Annotation und die Vielfalt der Arten von Formularen. So werden für Formulare bzgl. der Qualitätssicherung von medizinischen Geräten gute Resultate erzielt, wohingegen die Qualität der Annotationen für Formulare bzgl. der Ein- und Ausschlusskriterien von Studien ausbaufähig ist.

Der Aufbau dieser Arbeit ist wie folgt gegliedert. In Abschnitt 2 wird das Problem der Annotation von Formularen formal definiert. Der basale Workflow für die Identifikation der Annotationen ist in Abschnitt 3 erläutert. In Abschnitt 4 werden die zu realisierenden Erweiterungen für den definierten Workflow vorgestellt, um die Qualität der Annotationen zu verbessern und die Effizienz des Verfahrens zu erhöhen. In Abschnitt 5 wird konzeptionell die Architektur eines FMS für medizinische Formulare und ihre Annotationen vorgestellt. In Abschnitt 6 wird die Arbeit zusammengefasst.

2. PROBLEMDEFINITION

Das Ziel der semi-automatischen Annotation eines Formulars F ist die Bestimmung eines Annotations-Mappings $\mathcal{M}$ zwischen den Fragen $F = \{q_1, q_2, ...q_k\}$ des Formulars und den Konzepten $UMLS = \{cui_1, cui_2, ...cui_n\}$ des UMLS. Eine Annotation stellt eine Assoziation zwischen einer Frage und einem Konzept des UMLS dar, wobei eine Frage mit mehreren Konzepten annotiert sein kann. Dabei ist ein Konzept durch einen *Concept Unique Identifier* CUI eindeutig identifizierbar und wird durch Attribute wie z.B. einen Namen oder Synonyme beschrieben. Ein Annotations-Mapping $\mathcal{M}_{F,UMLS}$ ist formal definiert als:
$\mathcal{M}_{F,UMLS} = \{(q, cui, sim) | q \in F \wedge cui \in UMLS \wedge sim \in [0, 1]\}$. Dabei ist *sim* ein numerischer Wert, der die Ähnlichkeit zwischen einer Frage q und einem Konzept *cui* repräsentiert.

3. BASIS-WORKFLOW

Unser Ansatz für die Identifikation von Korrespondenzen basiert auf der Berechnung von Stringähnlichkeitsmaßen zwischen den Fragen der *Items* und den Attributen, wie z.B. den Namen und den Synonymen der Konzepte. Der generelle Workflow für die automatische Annotation ist in Abb. 2 dargestellt. Die Eingabe ist eine Menge von Formularen $\{F_1, F_2, ..., F_n\}$, das $UMLS$ und die Ausgabe ist eine Menge von Annotations-mappings $\{\mathcal{M}_{F_1,UMLS}, \mathcal{M}_{F_2,UMLS}, ..., \mathcal{M}_{F_n,UMLS}\}$. Zu Beginn werden im *Preprocessing* Schritt die Fragen bzw. Attribute der Konzepte normalisiert. Konkret, werden alle nicht relevanten Wörter entfernt, dazu gehören Präpositionen, Verben und Stoppwörter, die mithilfe eines Part-of-speech Taggers ermittelt werden. Des Weiteren werden alle Tokens klein geschrieben. Um eine effiziente Mapping-Generierung zu ermöglichen werden alle Tokens und Trigramme der Attribute der Fragen eines Formulars bzw. eines Konzepts enkodiert.

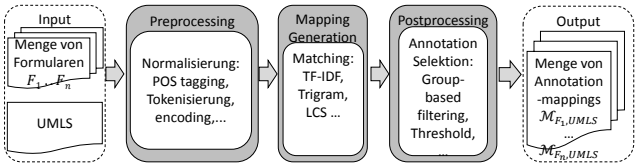


Figure 2: Annotations-Workflow

Im Schritt *Mapping-Generation* wird eine Menge von Tupeln der Form (q, cui, sim) durch den Vergleich der Fragen mit den Attributen der UMLS Konzepten generiert. Der Vergleich kann durch verschiedene Match-Verfahren realisiert werden wie z.B. Trigramm, TF/IDF oder Longest Common Substring (LCS). Bei einem naiven Ansatz wird das kartesische Produkt bzgl. der Menge der Fragen und der Menge der Konzepte berechnet, jedoch kann durch Pruning-Techniken oder partitionsbasiertes Matching die Anzahl der durchzuführenden Vergleiche reduziert werden [10].

In der *Postprocessing* Phase wird das Mapping durch die Anwendung von Aggregations- und Selektionsstrategien generiert. Im Allgemeinen wird eine Mindestähnlichkeit δ für eine Korrespondenz gefordert, damit diese als korrekt angesehen wird. Da es sich um einen semi-automatischen Prozess handelt, werden die identifizierten Annotationen durch einen Experten verifiziert.

4. ANSÄTZE ZUR ERWEITERUNG DES BASIS-WORKFLOWS

Aufgrund der Besonderheiten bzgl. der Annotation von Formularen, werden im Folgenden die Schwierigkeiten bzgl. des Annotationsprozesses beschrieben und mögliche Lösungsansätze erläutert.

Vorkommen natürlicher Sprache Im Gegensatz zu Ontologien, bei denen die Attribute der Konzepte in einer kompakten Repräsentation dargestellt sind, enthält eine Frage einen hohen Anteil an Freitext.

Ein möglicher Ansatz ist die Identifikation der Schlüsselwörter, die die Frage charakterisieren und ein Konzept des UMLS darstellen. Aufgrund des Vorkommens von Synonymen innerhalb einer Frage, die nicht in einem Konzept des UMLS erfasst sind, ist es nicht möglich durch Stringähnlichkeiten solche Korrespondenzen zu identifizieren. Ein Ansatz ist die Verwendung eines Synonymwörterbuchs, das es er-

laubt alle Tokens innerhalb einer Frage und eines Konzepts durch einen Identifier zu ersetzen. Mithilfe des Identifiers werden Synonyme als gleich angesehen, obwohl die Stringähnlichkeit gering ist. Das Synonymwörterbuch kann entweder durch externe Web-services generiert werden oder durch bereits verifizierte Annotationen erstellt werden.

Komplexe Mappings: Im Gegensatz zu Ontologie-Mappings, die im Allgemeinen aus 1:1 Korrespondenzen zwischen den Konzepten bestehen, werden komplexe Fragen in Formularen durch mehrere Konzepte inhaltlich beschrieben. Um solche komplexen Korrespondenzen zu identifizieren, sind die herkömmlichen Selektionsstrategien wie z.B. die Selektion der Korrespondenz mit der maximalen Ähnlichkeit oder Top-k nicht ausreichend. Für die Bestimmung dieser Korrespondenzen sind komplexe Selektionsstrategien oder entsprechende Vorverarbeitungsschritte sinnvoll. Im Folgenden wird eine Selektionsstrategie und eine mögliche Vorverarbeitung erläutert.

Bei der komplexen Selektionsstrategie werden die Korrespondenzen eines berechneten Mappings gefiltert, indem alle berechneten korrespondierenden Konzepte zu einer Frage bzgl. ihrer Ähnlichkeit gruppiert werden und pro Gruppe das Konzept als korrekt angesehen wird, welches die höchste Ähnlichkeit *sim* zu der Frage aufweist. Alle anderen Konzepte der Gruppe werden aus dem Mapping $\mathcal{M}_{F,UMLS}$ entfernt. Dieser Ansatz ist bereits realisiert und in der eingereichten Publikation vorgestellt.

Des Weiteren sind komplexe Korrespondenzen identifizierbar, wenn die Frage bzgl. ihres Inhalts separiert wird. Eine Wortgruppe oder Teilmenge der Frage repräsentiert dabei eine semantische Einheit und wird zu einem Konzept gematcht. Die Identifikation solcher Gruppen ist beispielsweise durch Named Entity Recognition (NER) Verfahren realisierbar oder durch eine statistische Erhebung von häufig auftretenden Kookkurrenzen innerhalb einer Menge von Formularen.

Größe der Datenquellen Das UMLS umfasst ~ 2.8 Mio. Konzepte, wohingegen ein Formular im Schnitt 50 Fragen enthält. Wenn man 100 Formulare annotiert, bedeutet dies, dass 14 Milliarden Vergleiche durchzuführen sind. Um einen effizienten automatischen Annotationsprozess zu realisieren ist es deshalb notwendig unnötige Vergleiche zu vermeiden.

Ein Ansatz zur Reduktion der Vergleiche ist die Verwendung von Bitlisten. Dabei wird das UMLS in Partitionen aufgeteilt. Bei der Partitionierung werden alle Konzepte bzgl. ihres Namens sortiert und einer Partition mit einer fixen Partitionsgröße (z.B. 100) zugeordnet. Alle Trigramme des Namens und der Synonyme eines Konzepts werden mittels einer Hashfunktion h auf eine Bitposition einer Bitliste der Länge l abgebildet. Die Trigramme werden tokenweise für das jeweilige Attribut erzeugt. Eine Bitlistenlänge $l = 27000$ ist ausreichend, wenn man ausschließlich kleingeschriebene Buchstaben berücksichtigt. Alle Bitlisten der Konzepte einer Partition werden durch die OR-Bitoperation zu einer Bitliste aggregiert. Die resultierende Bitliste ist ein Repräsentant der jeweiligen Partition. Ein Vergleich zwischen einer Frage und den Konzepten einer Partition wird durchgeführt, wenn der Bitlistenvektor der Frage, der ebenfalls durch die Hashfunktion h erstellt wird, eine geforderte relative Überlappung $min\_overlap[0, 1]$ erzielt. Die Berechnung der Überlappung entspricht der AND-Bitoperation. Die relative Überlappung *rel_overlap* ist der Quotient aus der Anzahl der Überlappung und der Anzahl der gesetzten Bits der Frage. Somit wird die Anzahl der Vergleiche für eine Frage auf die Anzahl der Konzepte beschränkt, die eine Mindestähnlichkeit bzgl. der Trigramme aufweisen.

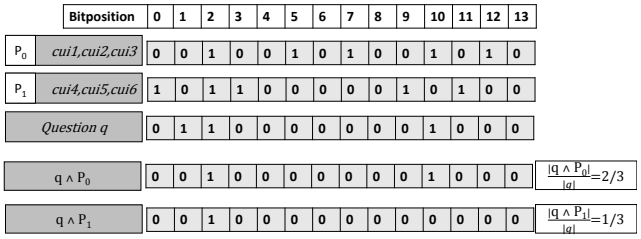


Figure 3: Beispiel für die Reduktion der Vergleiche mittels Partitionierung und der Repräsentation als Bitliste von Trigrammen

Ein Beispiel ist in Abb. 3 dargestellt, dabei wird die Menge der Konzepte $UMLS\_example = \{cui1, cui2, ...cui6\}$ und eine Frage q betrachtet. Die gegebene Menge wird auf die Partitionen P_0 und P_1 aufgeteilt. Dabei bilden die Trigramme der Konzepte $cui1$, $cui2$ und $cui3$ mittels einer Hashfunktion h auf die Bitpositionen $2, 5, 7, 10, 12$ ab. Analog wird der Bitlistenvektor für die Partition P_1 und die Frage q erstellt. Die relative Überlappung der Bitlisten der Frage q und der Partition P_0 ist $\frac{2}{3}$ und für P_1 $\frac{1}{3}$. Bei einer geforderten relativen Überlappung $min\_overlap = 0.5$ wird der Vergleich zwischen der Frage und den Konzepten $cui4$, $cui5$ und $cui6$ nicht durchgeführt, da die relative Überlappung $rel\_overlap = \frac{1}{3}$ ist.

Jedoch ist die Reduktion abhängig von der Effektivität der Partitionierung, so dass im ungünstigen Fall die Konzepte aller Partitionen verglichen werden müssen, wenn die Bitlisten eine hohe Überlappung untereinander aufweisen. Aufgrund dessen, ist eine qualitative Partitionierung bzgl. der Ähnlichkeit der Konzepte essentiell. Eine qualitativhochwertige berechnete Partitionierung ist unabhängig von den zu annotierenden Formularen, so dass diese für eine Vielzahl von Formularen einsetzbar ist.

5. ARCHITEKTUR EINES FORMULAR MANAGEMENT SYSTEMS (FMS)

Es ist geplant, ein Managementsystem zu realisieren, das die Formulare, Ontologien und die dazugehörigen Annotationen verwaltet. Das FMS soll die Möglichkeit bieten Formulare strukturiert zu suchen, ermittelte Annotationen zu verifizieren und neue Formulare zu annotieren. Das Managementsystem soll Wissenschaftlern die Möglichkeit bieten, effizient Formulare zu analysieren und passende Formulare wiederzuverwenden. Die Architektur umfasst eine Datenhaltungsschicht, eine Service-Schicht und eine Frontend-Schicht in Form einer Webanwendung(siehe Abb. 4).

Die Datenhaltungsschicht umfasst die Persistierung der Formulare, Ontologien und der berechneten sowie vorgeschla-

genen Annotationen durch eine relationale Datenbank. Die Service-Schicht umfasst folgende Module: *Import, Annotating, Search, Clustering* und *Verification.*

- **Import** Mithilfe des *Import* Moduls sollen Formulare in das Repository eingepflegt werden, so dass eine effiziente Suche bzw. Annotation möglich ist.
- **Annotating** Das *Annotating-Modul* ermöglicht die Annotation der Formulare des Repositories mit gewählten Ontologien. Des Weiteren sollen bereits annotierte Fragen verwendet werden, um unbekannte Fragen zu annotieren. Diesbezüglich ist ein Suchverfahren innerhalb des *Search-Moduls* notwendig, welches ähnliche Fragen oder Fragmente zu einer gegebenen Frage bzw. Fragments identifiziert. Die Annotationen der identifizierten Fragen sind mit hoher Wahrscheinlichkeit ebenfalls Annotationen für die gegebene Frage. Mithilfe der Wiederverwendung bereits existierender Annotationen wird der Vergleich mit dem kompletten UMLS vermieden.
- **Search** Um eine strukturierte Suche nach ähnlichen Formularen oder Fragen zu ermöglichen, umfasst das *Search-modul* eine Komponente, die basierend auf den Annotationen und der Eingabe einer Menge von Schlüsselwörtern eine explorative Suche nach den gewünschten Formularen bzw. Fragen ermöglicht. Des Weiteren soll dieses Modul eine Komponente umfassen, die eine effiziente Suche nach ähnlichen Fragen ermöglicht. Ein naiver Ansatz wäre die Erstellung einer invertierten Liste bzgl. der Token oder Wortgruppen einer Frage, um für eine unbekannte Frage, die ähnlichsten Fragen zu ermitteln.
- **Clustering** Des Weiteren kann die Effizienz der Suche durch eine Clusterung der Formulare bzw. Fragen erhöht werden. In diesem Modul sollen Clustering-Verfahren bereitgestellt werden, die basierend auf den Annotationen eine Gruppierung der Formulare und Fragen ermöglichen.
- **Verification** Da ein automatisches Verfahren keine vollständige Korrektheit gewährleisten kann, soll dieses Modul die Bewertung von Experten in den Qualitätssicherungsprozess bzgl. der Annotationen mit einbeziehen. Ein Experte soll in der Lage sein berechnete Annotationen zu bewerten oder zu ergänzen. Somit soll eine stetige Verbesserung der Qualität der Annotationen im System erzielt werden. Des Weiteren soll mithilfe der verifizierten Annotationen die Effektivität und Effizienz des Annotationsprozesses mittels der Wiederverwendung erhöht werden.

Die Frontend-Schicht wird durch eine Webanwendung repräsentiert, so dass der Anwender die Möglichkeit hat neue Formulare zu importieren, ähnliche Formulare oder Teilfragmente mithilfe einer explorativen Suchfunktion zu ermitteln. Der Anwender soll durch die Eingabe eines Suchterms die Möglichkeit haben, die Menge der Formulare mittels der Annotationen weiter einzugrenzen. Ein Ansatz für eine explorative Suche mittels einer Tag-Cloud ist in eTACTS [9] realisiert. Des Weiteren soll eine Sicht für die Verifikation

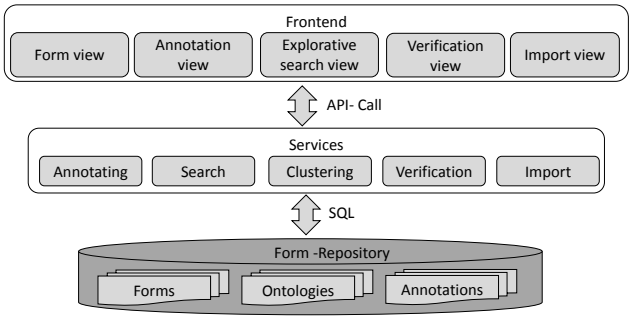


Figure 4: Architektur eines FMS

bereitgestellt werden, die Experten erlaubt einzelne Annotationen zu bewerten.

6. ZUSAMMENFASSUNG

Annotationen sind für die Beschreibung und einheitliche Repräsentation von Formularen essentiell. Durch die Verwendung von Annotationen wird der Datenaustausch, die Integration von Daten der zugrundeliegenden Formulare und die Suche vereinfacht. Um einen effektiven und effizienten Annotationsprozess zu realisieren, sind die bisherigen Methoden des Ontologie-Matching nicht ausreichend. In dieser Arbeit wurde der generelle Workflow für die semi-automatische Annotation vorgestellt sowie Lösungsansätze präsentiert, die die Besonderheiten der Annotation von Formularen behandeln. Um den Nutzen der Allgemeinheit zur Verfügung zu stellen, wurde konzeptionell die Architektur eines Formular Management Systems dargestellt, welches die Möglichkeit bietet neben der Annotation, Formulare oder Fragen basierend auf den Annotationen zu suchen oder zu analysieren. Aufgrund des automatischen Annotationprozesses soll im Gegensatz zur MDM-Plattform die Vielzahl der Formulare annotiert sein. Da jedoch ein automatisches Verfahren keine vollständige Korrektheit gewährleisten kann, soll mithilfe einer Verification-Komponente ein Expertenkonsortium für die Verifikation mit einbezogen werden.

7. REFERENCES

[1] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl 1):D267–D270, 2004.

[2] B. Breil, J. Kenneweg, F. Fritz, et al. Multilingual medical data models in ODM format–a novel form-based approach to semantic interoperability between routine health-care and clinical research. *Appl Clin Inf*, 3:276–289, 2012.

[3] K. Donnelly. SNOMED-CT: The Advanced Terminology and Coding System for eHealth. *Studies in Health Technology and Informatics–Medical and Care Compunetics 3*, 121:279–290, 2006.

[4] M. Dugas. Missing Semantic Annotation in Databases. The Root Cause for Data Integration and Migration Problems in Information Systems. *Methods of Information in Medicine*, 53(6):516–517, 2014.

[5] P. Indyk and R. Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98,

pages 604–613, New York, NY, USA, 1998. ACM.

[6] T. Kirsten, A. Gross, M. Hartung, and E. Rahm. GOMMA: a component-based infrastructure for managing and analyzing life science ontologies and their evolution. *Journal of Biomedical Semantics*, 2(6), 2011.

[7] A. J. Kirtane, A. Gupta, S. Iyengar, J. W. Moses, M. B. Leon, R. Applegate, B. Brodie, E. Hannan, K. Harjai, L. O. Jensen, et al. Safety and efficacy of drug-eluting and bare metal stents comprehensive meta-analysis of randomized trials and observational studies. *Circulation*, 119(25):3198–3206, 2009.

[8] H. J. Lowe and G. O. Barnett. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Journal of the American Medical Association (JAMA)*, 271(14):1103–1108, 1994.

[9] R. Miotto, S. Jiang, and C. Weng. eTACTS: A method for dynamically filtering clinical trial search results. *Journal of Biomedical Informatics*, 46(6):1060–1067, 2013.

[10] E. Rahm. Towards Large-Scale Schema and Ontology Matching. In Z. Bellahsene, A. Bonifati, and E. Rahm, editors, *Schema Matching and Mapping*, Data-Centric Systems and Applications, pages 3–27. Springer Berlin Heidelberg, 2011.

[11] P. Shvaiko and J. Euzenat. A survey of schema-based matching approaches. In *Journal on Data Semantics IV*, pages 146–171. Springer, 2005.

[12] C. Xiao, W. Wang, X. Lin, and J. X. Yu. Efficient similarity joins for near duplicate detection. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 131–140, New York, NY, USA, 2008. ACM.