Fachhochschule Köln
Cologne University of Applied Sciences

Semi-automatische Verschlagwortung zur Integration externer semantischer Inhalte innerhalb einer medizinischen Kooperationsplattform

Masterarbeit

ausgearbeitet von

Zeljko Carevic

zur Erlangung des akademischen Grades

Master of Science

vorgelegt an der

Fachhochschule Köln
Campus Gummersbach

im Studiengang

Medieninformatik

Erster Prüfer:	Prof. Dr. Kristian Fischer Fachhochschule Köln
Zweiter Prüfer:	Prof. Wolfgang Prinz, PhD Fraunhofer-Institut für Angewandte Informationstechnik FIT

Wuppertal, April 2012

Adressen: Zeljko Carevic
Beeker Winkel 10
42113 Wuppertal
Z.Carevic@googlemail.com

Prof. Dr. Kristian Fischer
Fachhochschule Köln
Institut für Informatik
Steinmüllerallee 1
51643 Gummersbach
kristian.fischer@fh-koeln.de

Prof. Wolfgang Prinz, PhD
Fraunhofer-Institut für Angewandte Informationstechnik FIT
Schloss Birlinghoven
53754 Sankt Augustin
wolfgang.prinz@fit.fraunhofer.de

Inhaltsverzeichnis

1 Einleitung

Die heutige Informationsgesellschaft produziert soviel Wissen, dass man von einer Informationsexplosion sprechen kann. In vielen Bereichen, darunter auch die Medizin, zählt daher die Informationsbeschaffung zu den wesentlichen Kompetenzen. Die Menge an Informationen die es dabei zu betrachten gilt, führt jedoch nicht selten zu einer Desorientierung bei der Informationsbeschaffung. Die Qualität der medizinischen Versorgung hängt neben Aspekten wie der Ausbildung von medizinischem Personal, der technischen Ausstattung der Einrichtung oder organisatorischen Aspekten auch wesentlich von der Verfügbarkeit von Informationen ab. Unterstützend können Informationssysteme eingesetzt werden die in der Lage sind die Verfügbarkeit, Vollständigkeit, sowie Fehlerfreiheit von Informationen zu verbessern, um auf diese Weise Ärzten und Pflegepersonal eine raschere Orientierung zu ermöglichen (Lenz et al., 2005).

Aus informationstechnischer Sicht existieren zahlreiche Forschungsbereiche die sich mit einer Optimierung der Informationsbeschaffung befassen. Zu nennen ist hier vor allem der Bereich des Information Retrieval (IR) mit dem Ziel, Nutzer bei der Suche nach Dokumenten, die einen Informationsbedarf erfüllen zu unterstützen (Baeza-Yates und Ribeiro-Neto, 1999). IR-Systeme kennzeichnen sich dadurch, dass Wissen unsicher ist was aus der begrenzten Repräsentation von dessen Semantik resultiert (z.B. bei Texten oder multimedialen Dokumenten) (Fuhr, 1996).

Betrachtet man beispielsweise Informationen im WWW wird erkennbar, dass diese in der Regel auf menschliche Nutzer als Konsumenten optimiert sind. Eine maschinelle Verarbeitung webbasierter Inhalte gestaltet sich oft als schwierig da die Semantik der Inhalte für Maschinen nicht erkennbar ist. Suchmaschinen, deren Algorithmen auf Basis einer Volltextsuche arbeiten sind nicht in der Lage homographe[1] Wörter zu unterscheiden, was zu unerwünschten Ergebnissen führen kann.

Ein Ansatz Wissen maschinell interpertierbar zu machen, wird durch das Konzept des *Semantic Web* beschrieben. Das Semantic Web ist als Erweiterung des heutigen WWW zu verstehen, das Informationen unter Verwendung offener Standards in strukturierter

[1] Orthographisch gleiche Wörter mit unterschiedlicher Bedeutung (Mehrdeutigkeit)

Form aufbereitet. Auf diese Weise wird es möglich die Kooperation zwischen Menschen und Maschinen zu optimieren (Berners-Lee et al., 2001). Ist einem System die Semantik der Inhalte bekannt führt dies zu einer signifikanten Verbesserung der Informationsbeschaffung und einer Reduzierung von Unsicherheit.

Zur strukturierten Aufbereitung von Informationen haben sich Standards entwickelt - bekannt als Semantic Web Technologien. Mittels semantischer Technologien wird ein offener Standard für die Wissensrepräsentation geschaffen der einen Informationsaustausch zwischen verschiedenen Anwendungen und Plattformen realisierbar macht (Hitzler et al., 2008).

Insbesondere die Propagierung der Linked Data Prinzipien durch Tim Berners-Lee (2006) führte zu einer Entwicklung dezentraler Wissensbasen im Web, die mittels semantischer Technologien aufbereitet sind. So umfasst das *Linking Open Data* Projekt zur Zeit[2] bereits mehr als 290 semantisch aufbereitete Wissensbasen mit knapp 32 Mrd. Triples[3]. Der Bereich der Lebenswissenschaften wird durch 41 Wissensbasen repräsentiert und umfasst Informationen zu klinischen Studien, Medikamenten, Krankheiten etc. die mittels semantischer Relationen verknüpft sind (Bizer et al., 2011). Durch Verwendung semantischer Relationen wird es möglich, dass Wissen miteinander verknüpft wird. Auf diese Weise lässt sich strukturiertes Wissen über eine Krankheit mit strukturiertem Wissen über bspw. klinische Studien oder Medikamente verbinden.

Im Wesentlichen dient das Linking Open Data Projekt der Identifikation lizenzfreier Datenbanken im Web, die nach den Linked Data Prinzipien konvertiert und veröffentlicht werden (Bizer et al., 2009). Insbesondere bei Inhalten aus den Lebenswissenschaften wird die verwendete Terminologie aus den Quelldaten übernommen. Durch eine Reihe von Standardisierungsbemühungen haben sich innerhalb der Biomedizin jedoch zahlreiche Begriffssysteme entwickelt die von verschiedenen Institutionen gepflegt und verwendet werden, mit der Konsequenz einer inkonsistenten Begriffsdefinition und dem Fehlen einer allgemein akzeptierten Terminologie. Dies spiegelt sich auch in Wissensbasen der Linked Data Cloud wieder. So verwenden unterschiedliche Wissensbasen in der Linked Data Cloud auch unterschiedliche Bezeichner für ihre Ressourcen.

Um das Fehlen konsistenter Begriffsdefinitionen in der Biomedizin zu kompensieren wurde das Unified Medical Language System (UMLS) entwickelt, das mehr als 100 Begriffssysteme mit 10 Mio. Bezeichnungen zu mehr als 2,5 Mio. Konzepten integriert und redundante Terme auf einen gemeinsamen Deskriptor zusammenfasst (Bodenrei-

[2]Stand 09/2011
[3]Als Triple bezeichnet man Aussagen in Form von Subjekt, Prädikat, Objekt

der, 2004),(UMLS Statistics 2011). Dazu zählt auch das zur Indexierung medizinischer Publikationen verwendete kontrollierte Vokabular MeSH. Durch Überschneidungen in der verwendeten Terminologie in der Linked Data Cloud und dem UMLS bietet sich hier Potential einer konsistenten Integration von Informationen. Da Wissen in der Linked Data Cloud in strukturierter Form aufbereitet ist lässt sich zu zahlreichen Ressourcen die verwendete Terminologie identifizieren. DBPedia bspw. verwendet bei der Aufbereitung medizinischer Inhalte häufig das kontrollierte Vokabular MeSH und kennzeichnet dies entsprechend. Auf diese Weise wird es möglich trotz der Begriffsvielfalt in der Biomedizin passende Ressourcen in der Linked Data Cloud zu identifizieren.

Zwar bietet ein kontrolliertes Vokabular bei der Informationsbeschaffung zahlreiche Vorteile gegenüber einer Freitextsuche doch fällt Nutzern das Abbilden eines Informationsbedarfs auf die verwendete Terminologie oftmals schwer (Gault et al., 2002). Hier wird eine Systemunterstützung geschaffen, die den Abbildungsprozess automatisiert indem eine automatische Verschlagwortung textbasierter Inhalte unter Verwendung eines kontrollierten Vokabulars vorgenommen wird. Diese werden dann unter Verwendung einer einheitlichen Terminologie mit externen Informationen aus der Linked Data Cloud und PubMed angereichert.

1.1 Fragestellungen

Die vorliegende Arbeit beschäftigt sich mit der Integration von externen semantischen Inhalten auf Basis eines medizinischen Begriffssystems. Die zugrundeliegende Annahme ist, dass die Verwendung einer einheitlichen Terminologie auf Seiten des Anfragesystems und der Wissensbasis zu qualitativ hochwertigen Ergebnissen führt. Um dies zu erreichen muss auf Seiten des Anfragesystems eine Abbildung natürlicher Sprache auf die verwendete Terminologie gewährleistet werden. Dies geschieht auf Basis einer (semi-)automatischen Verschlagwortung textbasierter Inhalte. Im Wesentlichen lassen sich folgende Fragestellungen festhalten:

Automatische Verschlagwortung textbasierter Inhalte
Kann eine automatische Verschlagwortung textbasierter Inhalte auf Basis eines Begriffssystems optimiert werden?

Der zentrale Aspekt der vorliegenden Arbeit ist die (semi-)automatische Verschlagwortung textbasierter Inhalte auf Basis eines medizinischen Begriffssystems. Zu diesem Zweck wird der aktuelle Stand der Forschung betrachtet. Es werden eine Reihe von *Tokenizern* verglichen um zu erfahren welche Algorithmen sich zur Ermittlung von Wortgrenzen eignen. Speziell wird betrachtet, wie die Ermittlung von Wortgrenzen in einer domänenspezifischen Umgebung eingesetzt werden kann. Auf Basis von identifizierten Token in einem Text werden die Auswirkungen des *Stemming* und *POS-Tagging* auf die Gesamtmenge der zu analysierenden Inhalte beobachtet.

Abschließend wird evaluiert wie ein kontrolliertes Vokabular die Präzision bei der Verschlagwortung erhöhen kann. Dies geschieht unter der Annahme dass domänenspezifische Inhalte auch innerhalb eines domänenspezifischen Begriffssystems definiert sind. Zu diesem Zweck wird ein allgemeines Prozessmodell entwickelt anhand dessen eine Verschlagwortung vorgenommen wird.

Integration externer Inhalte
Inwieweit kann die Nutzung einer einheitlichen Terminologie zwischen Anfragesystem und Wissensbasis den Prozess der Informationsbeschaffung unterstützen?

Zu diesem Zweck wird in einer ersten Phase ermittelt welche Wissensbasen aus der medizinischen Domäne in der Linked Data Cloud zur Verfügung stehen. Aufbauend auf den Ergebnissen werden Informationen aus verschiedenen dezentralen Wissensbasen exemplarisch integriert. Der Fokus der Betrachtung liegt dabei auf der verwendeten Terminologie sowie der Nutzung von Semantic Web Technologien.

Neben Informationen aus der Linked Data Cloud erfolgt eine Suche nach medizinischer Literatur in PubMed. Wie auch in der Linked Data Cloud erfolgt die Integration unter Verwendung einer einheitlichen Terminologie. Eine weitere Fragestellung ist, wie Informationen aus insgesamt 21. Mio Aufsatzzitaten in PubMed sinnvoll integriert werden können. Dabei wird ermittelt welche Mechanismen eingesetzt werden können um die Präzision der Ergebnisse zu optimieren.

Eignung medizinischer Begriffssystem

Welche medizinischen Begriffssysteme existieren und wie eignen sich diese als zugrungeliegendes Vokabular für die automatische Verschlagwortung und Integration semantischer Inhalte?
Der Fokus liegt dabei speziell auf einer Bewertung der Reichhaltigkeit von Begriffssystemen, wobei insbesondere der Detaillierungsgrad von Interesse ist. Handelt es sich um ein spezifisches oder allgemeines Begriffssystem und eignet sich dieses auch dafür bestimmte Teilaspekte der Medizin, wie bspw. die Chirurige oder die Anästhesie, in einer ausreichenden Tiefe zu beschreiben?

1.2 Projektgegenstand

Die vorliegende Arbeit ensteht im Rahmen eines interdisziplinären Forschungsprojekts, das in Zusammenarbeit mit verschiedenen Universitäten und Unternehmen entwickelt wird. Das Ziel ist die Entwicklung einer innovativen Web 2.0 Weiterbildungsplattform für die Chirurgie, namens *SurgeryTube*.

Im Fokus steht besonders der kooperative Austausch einer breiten und wachsenden Zahl von qualifizierungsgeeigneten Videos für Chirurgen (Mildner, 2009).

Die automatische Verschlagwortung textbasierter Inhalte sowie die daraus resultierende Integration externer Informationen am Beispiel von SurgeryTube lässt sich durch folgendes Szenario verdeutlichen:

> Dr. Riedel ist seit 25 Jahren Chirurg an der Universitätsklinik Lübeck. Zur Zeit bereitet sich Dr. Riedel auf eine, per Stream in Echtzeit übertragene Operation (Live-Operation) an der Leber vor, die in einer Woche Studenten der Universität Lübeck gezeigt werden soll. Dr. Riedel möchte den Studenten nicht nur technische Fähigkeiten vermitteln sondern auch benötigtes Informationsmaterial und Grundlagenwissen an die Hand geben, um die Studenten optimal auf den Eingriff vorzubereiten. Zu diesem Zweck meldet sich Dr. Riedel bei SurgeryTube an und verfasst eine kurze Beschreibung zu dem geplanten Eingriff. Dazu zählt neben allgemeinen Informationen über den geplanten Eingriff auch eine Beschreibung der Diagnose. Dr. Riedel speichert seinen Beitrag in SurgeryTube und das System

stellt ihm eine Reihe von möglichen Schlagwörtern zu dem Text vor, die geeignet sind das behandelte Thema zu beschreiben. Nach kurzer Betrachtung der Schlagwörter wählt Dr. Riedel sechs Stück aus und verlässt SurgeryTube.

Christian S. studiert im fünften Semester Medizin an der Universität Witten/Herdecke. Für morgen steht eine Live-Operation von Dr. Riedel bevor, der Christian via Live-Stream beiwohnen möchte. Zur Vorbereitung meldet er sich bei SurgeryTube an und liest sich die von Dr. Riedel verfasste Beschreibung des Eingriffs vor. Dabei muss er erkennen, dass ihm die beschriebene Diagnose nicht bekannt ist. Um sich mit Hintergrundinformationen zu versorgen wählt Christian eines der Schlagwörter aus, die von Dr. Riedel vergeben wurden. Das System präsentiert Christian nützliches Hintergrundwissen. Dazu zählen wissenschaftliche Artikel aus PubMed, eine Begriffsdefinition sowie Informationen zu Genprodukten und Medikamenten.

1.3 Aufbau der Arbeit

Im Folgenden werden zunächst einige Grundlagen rekapituliert, die für ein besseres Verständnis der Folgeinhalte notwendig sind. Dazu zählen sowohl Konzepte der Informationsbeschaffung als auch Konzepte der Wissensmodellierung. In Kapitel 3 wird der aktuelle Stand der Technik hinsichtlich der automatischen Verschlagwortung und Aggregation medizinscher Inhalte erläutert, die losgelöst voneinander betrachtet werden.

In Abschnitt 4 werden zwei medizinische Begriffssysteme, das *Medical Subject Heading* (MeSH) und das *Unified Medical Language System* (UMLS) vorgestellt. Diese werden hinsichtlich ihrer Eignung als zugrundeliegendes Begriffssystem für die weitere Verarbeitung verglichen. Der Schwerpunkt der vorliegenden Arbeit liegt auf der automatischen Verschlagwortung von Inhalten. Dies wird in Kapitel 5 erläutert und setzt sich aus den Abschnitten Tokenizing und Filterung zusammen auf dessen Basis innerhalb einer Merkmalsanalyse ein Bewertungskriterium für die Relevanz eines Token vorgestellt werden. Abschließend folgt eine Bewertung der erzielten Ergebnisse anhand von Beispieltexten.

Aufbauend auf den Ergebnissen der automatischen Verschlagwortung erfolgt in Kapitel 6 eine Integration externer Inhalte. Dazu zählt die Integration medizinischer Publikationen zu ausgewählten Schlagwörtern, sowie eine Integration von Inhalten aus der Linked Data Cloud.

2 Grundlagen

Zunächst werden Konzepte der Informationsbeschaffung rekapituliert. Davon losgelöst werden in Abschnitt 2.2 eine Reihe von Ansätzen zur Wissensmodellierung präsentiert. Abschliessend folgt eine kurze Einführung zur bibliographischen Suche mittels PubMed.

2.1 Konzepte der Informationsbeschaffung

Im Wesentlichen kann die hier vorliegende Zielsetzung als Extraktion relevanter Informationen in schwach strukturierten, textbasierten Daten definiert werden. Textbasierte Inhalte zeichnen sich dadurch aus, dass sie in der Regel keiner festen Struktur folgen und von Maschinen lediglich als Reihe von Zeichen bzw. Zeichenketten interpretiert werden. Dennoch findet sich auch in Texten zumeist eine aus der Grammatik resultierende implizite Struktur. Daneben lassen sich auch einige wenige explizite Strukturen in Texten identifizieren die bspw. aus dem Titel oder aus einem Absatz resultieren (Hippner und Rentzmann, 2006). Im Folgenden werden eine Reihe von Ansätzen präsentiert, die sich mit der Extraktion von Informationen aus strukturierten und unstrukturierten Daten befassen.

Information Retrieval (IR) ist ein Forschungsbereich mit dem Ziel Nutzer bei der Suche nach Dokumenten, die einen Bedarf nach Information erfüllen zu unterstützen (Baeza-Yates und Ribeiro-Neto, 1999). IR kann als Frage - Antwort System bezeichnet werden, dass zu einer gegebenen Anfrage vorhandenes Wissen aus einem IR-System extrahiert. Die Gesellschaft für Informatik (GI) kennzeichnet IR-Systeme dadurch, dass Anfragen vage und Wissen unsicher ist, was zu der Notwendigkeit einer Bewertung der Antworten führt (Fuhr, 1996). In diesem Zusammenhang ist das primäre Qualitätsmerkmal von IR-Systemen die Relevanz der Ergebnisse. Ein typisches Beispiel für IR-Systeme sind Suchmaschinen bei denen zu vagen Formulierungen Informationen präsentieren werden, die geeignet sind ein Informationsbedürfnis zu befriedigen. Die Form in der die Daten eines IR-Systems vorliegen ist nicht beschränkt, jedoch liegt der Fokus

meist auf der Betrachtung von Text bzw. multimedialen Daten. Ein verallgemeinertes Modell der Anfrageverarbeitung findet sich in Abbildung 2.1. Eine Anfrage wird zunächst in ein Format transformiert, dass eine maschinelle Verarbeitung erleichtert, hier als die Transformation natürlicher Sprache in eine Anfragerepräsentation dargestellt. Aus Sicht des Informationssystems kann analog dazu eine Transformation von Dokumenten in eine Dokumentrepräsentation vorgenommen werden. Der eigentliche Prozess der Extraktion findet während des *matching* statt, bei dem auf Basis der Anfragerepräsentation relevante Informationen in der Dokumentrepräsentation ermittelt werden. Das Ergebnis wird in Form einer relevanzsortierten Liste den Benutzern präsentiert. Je nach IR-System folgt ein sogenanntes *Relevance Feedback*, das eine manuelle Gewichtung der Ergebnisse seitens der Nutzer erlaubt.

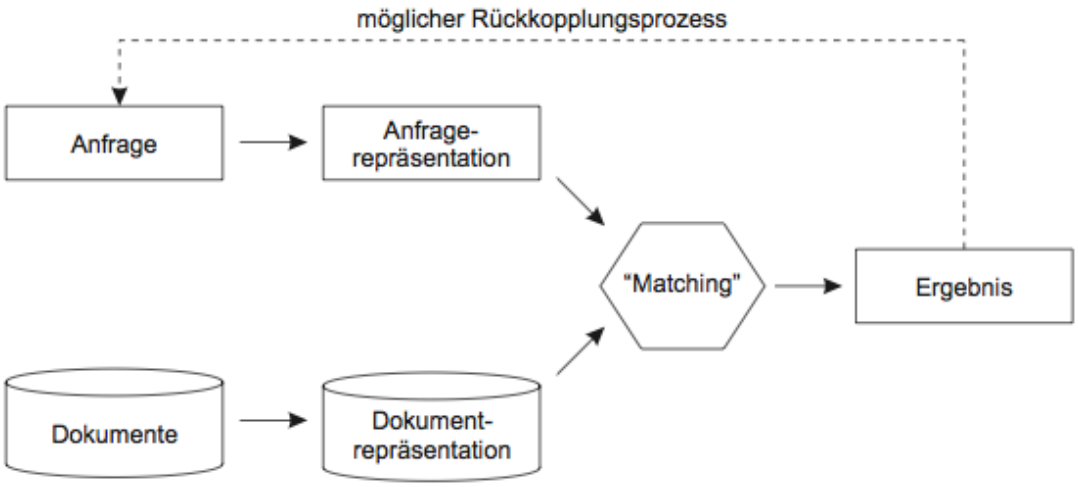


Abbildung 2.1: Verallgemeinertes Modell der Anfrageverarbeitung im IR. Quelle: (Henrich, 2008)

Knowledge Discovery in Data (KDD) ist ein Forschungsbereich, der darauf abzielt Nutzer bei der Analyse großer Datenmengen zu unterstützen, um so nützliches Wissen zu extrahieren (Reinartz, 1999). Nach Fayyad und Anderen ist KDD definiert als der "nichttriviale Prozess der Identifikation valider, neuer, potentiell nützlicher und verständlicher Muster in Daten" (Fayyad et al., 1996). Ein Beispiel für KDD Systeme findet sich im Marketing. Häufig wird das Kaufverhalten von Kunden analysiert um Muster zu erkennen und Vorhersagen treffen zu können. Das Datenvolumen solcher Systeme ist in der Regel so umfassend, dass eine manuelle Analyse nicht mehr unter vertretbarem Arbeitsaufwand durchgeführt werden kann. Daher kommen Techniken und Werkzeuge der KDD zum Einsatz die eine maschinelle Analyse der Daten ermöglichen, um auf diese Weise zu nützlichem Wissen zu gelangen. KDD ist gekennzeichnet

durch eine Reihe iterativer Prozesse. Dazu zählt die Auswahl geeigneter Datenbasen, Aufbereitung der Daten auf eine einheitliche Datenbasis, Reduzierung der Daten durch Transformation, Analyse (Data Mining) sowie die Interpretation/Evaluation.

Data Mining - Text Mining und KDD werden häufig synonym verwendet. Bei Data Mining handelt es sich jedoch lediglich um eine Phase innerhalb des KDD die nach Fayyad und Anderen (1996) definiert ist als:

> *"A step in the KDD process that consists of applying data analysis and discovery algorithms that produce a particular enumeration of patterns (or models) over the data."*

Data Mining kann als Anwendung von Algorithmen und Methoden verstanden werden die der Extraktion potentiell nützlicher Muster in der KDD dienen. Damit bildet Data Mining die zentrale Phase der Analyse in der KDD. Ein schon durch die Namensgebung eng mit dem Data Mining verbundener Ansatz ist das *Text Mining*. Dabei unterscheiden sich die beiden Ansätze darin, dass der Einsatz von Data Mining in der Regel auf Basis von strukturierten Daten, wie beispielsweise Datenbanken erfolgt, wohingegen das Text Mining vornehmlich im Bereich von unstrukturierten Daten verwendet wird. Eine Betrachtung von Text Mining aus Information Retrieval, Data Mining, methodischer und wissensorientierter Perspektive findet sich in Mehler und Wolff (2008).

Information Extraction bezeichnet die selektive Extraktion spezifizierter Informationstypen aus unstrukturierten Texten (Yangarber et al., 2000). Das Ziel ist es strukturierte Informationen aus unstrukturierten Daten (Text) zu gewinnen. Das Resultat einer IE kann bspw. eine tabellarische Repräsentation domänenrelevanter Eigenschaften sein. Generell kann die Aufgabe eines IE-Systems als Zusammenfassung eines uneingeschränkten Textes verstanden werden, wobei die Zusammenfassung unter Berücksichtigung eines zuvor spezifizierten Themas bzw. einer Domäne erfolgt (Cardie, 1997). In dem in Abbildung 2.2 dargestellten Beispiel einer IE werden Zeitungsartikel zum Thema Naturkatastrophen analysiert. Die Aufgabe einer IE besteht nun darin die tabellarisch dargestellte Maske mit Informationen aus dem Text anzureichern. Die in der Maske enthaltenen Attribute werden zuvor spezifiziert und dienen der Vorgabe thematisch relevanter Informationen. Entsprechend wird zu jedem thematisch zugehörigen Zeitungsartikel eine Extraktion der relevanten Informationen auf Basis der Maske durchgeführt wodurch ein reichhaltiges, strukturiertes Informationssystem entsteht.

> 4 Apr Dallas - Early last evening, a tornado swept through an area northwest of Dallas, causing extensive damage. Witnesses confirm that the twister occurred without warning at approximately 7:15 p.m. and destroyed two mobile homes. The Texaco station, at 102 Main Street, Farmers Branch, TX,was also severely damaged, but no injuries were reported. Total property damages are estimated to be $350,000.

Event	tornado
Date	4/3/97
Time	19:15
Location	Farmers Branch : "northwest of Dallas" : TX : USA
Damage	"mobile homes" (2)
	"Texaco station" (1)
Estimated Losses	$350,000
Injuries	none

Abbildung 2.2: Information Extraction am Beispiel einer Naturkatastrophe. Quelle: (Cardie, 1997)

2.1.1 Computerlinguistik

Die zuvor dargestellten Forschungsbereiche beschäftigen sich mit dem Transfer von Informationen aus großen Datenmengen. Unterschiede lassen sich jedoch insbesondere in der Zielsetzung und der verwendeten Datenbasis erkennen. So sind IR-Systeme in der Lage dem Nutzer Dokumente, die Informationen zu einer gegebenen Anfrage enthalten zu präsentieren. Daraus geht hervor, dass Informationen in einem IR-System grundsätzlich bereits vorhanden sein müssen. KDD hingegen zielt auf die Erkennung neuer Informationen aus Datenmengen ab, indem Muster erkannt werden und Rauschen eliminiert wird (Hearst, 1999). Abgrenzend dazu zielen IE-Systeme auf eine strukturierte, domänenspezifische Repräsentation unstrukturierter Daten ab, die im weiteren Verlauf als Datenbasis des Informationstransfers verwendet wird. Eine Gemeinsamkeit die jedoch allen diskutierten Bereichen zugrunde liegt ist, dass es zu einer Verarbeitung natürlicher Sprache kommt. Ein Bereich der sich mit der Verarbeitung von natürlicher Sprache beschäftigt ist die Computerlinguistik, die sowohl Text als auch gesprochene Sprache umfasst. Auch wenn unterschiedliche Auffassungen über Computerlinguistik existieren ist diese hier als Disziplin für die Entwicklung von Anwendungen zu verstehen, die eine linguistische Datenverarbeitung erlaubt (Carstensen et al., 2010). Häufig wird die Computerlingustik als Vorverarbeitungsprozess eingesetzt der eine genaue Analyse der Daten erst ermöglicht. Zu den Techniken die bei der Verarbeitung natürlicher Sprache

zum Einsatz kommen zählen:

Tokenizing Bei dem Vorgang des Tokenizing (*Tokenisieren* oder auch *Segmentieren*) handelt es sich um die Isolation wortähnlicher Einheiten (Token) in einem Text (Grefenstette und Pasi, 1994). Das Tokenizing bildet einen der zentralen Prozesse bei der Verarbeitung natürlicher Sprache da nahezu alle Folgeprozesse auf Basis von Token agieren.

Part-Of-Speech Tagger Bei sogenannten POS-Taggern handelt es sich um Algorithmen die Token eine bestimmten grammatikalischen Kategorie zuweisen. Auf diese Weise können Filtermechanismen angewandt werden die nur bestimmte grammatikalische Kategorien zulassen.

Stemming Als Stemming bezeichnet man die Rückführung morphologischer Varianten eines Begriffs auf seine Stammform. Ein Beispiel für eine morphologische Variante eines Begriffs ist "Monitor" ⇒ "Monitore" oder "weit" ⇒ "weiter".

2.1.2 Maschinelles Lernen

Bei maschinellem Lernen handelt es sich um einen der ältesten Forschungsbereiche der künstlichen Inteligenz der auf die 50er Jahre zurückgeht. Eine einheitlich akzeptierte Definition existiert jedoch nicht. In Beierle und Kern-Isberner (2006) werden jedoch eine Reihe von Definitionen diskutiert, deren zentrale Aspekte die Performanzsteigerung sowie die Konstruktion einer Repräsentation beinhalten. Menschen sind in der Lage bei einer ausgeführten Handlung zu einer Performanzsteigerung durch Erfahrungswerte zu gelangen. Maschinen hingegen sind strikte Ausführungseinheiten einer Prozedur die sich zwar durch enorme Effizienz auszeichnen, dabei jedoch keine Erfahrungswerte sammeln um zu einer Performanzsteigerung zu gelangen. Nach Michalski (1999) beschäftigt sich maschinelles Lernen mit der Entwicklung von Computerprogrammen, die in der Lage sind neues Wissen zu konstruieren, oder bereits vorhandenes Wissens zu verbessern. Hierzu werden Eingangsinformationen verwendet die Fakten, Beispiele, Beschreibungen etc. sein können und in der Regel von Menschen erzeugt werden.

Eine Aufgabe des maschinellen Lernens ist die **Klassifikation** bei dem sogenannte Merkmalsvektoren in eine endliche Menge von Klassen eingeteilt werden. Ein Merk-

malsvektor besteht aus N Merkmalen, die für eine Klassifikation relevant sind. Ein sehr einfaches Beispiel für eine Klassifikation ist in Tabelle 2.1 dargestellt. In dem Beispiel wird die Güteklasse von Äpfeln in Abhängigkeit der Merkmale Größe und Farbe (fiktiver Farbwert zwischen 0 für grün und 1 für rot) ermittelt. Entsprechend wird ein zweidimensionaler Merkmalsvektor erzeugt, der jeweils der Klasse A oder B zugeordnet wird. Bei realen Klassifikationsaufgaben werden in der Regel mehr als nur zweidimensionale Merkmalsvektoren verwendet.

Größe (cm)	Farbe	Klasse
6	0,1	B
8	0,7	A
7	0,9	A
4	0,2	B

Tabelle 2.1: Beispiel einer Klassifikation. Eigene Darstellung nach (Ertel, 2008)

Um eine maschinelle Klassifikation vorzunehmen kommen sogenannte Trainingsdaten zum Einsatz, die bereits vor-klassifiziert sind. Anhand dieser Trainingsdaten können nun Algorithmen des maschinellen Lernens trainiert werden, um auf Basis eines nichtklassifizierten Merkmalsvektors eine approximierte Klassifikation vorzunehmen. Dieser Vorgang wird auch als **überwachtes Lernen** (supervised learning) bezeichnet. Überwachtes Lernen zeichnet sich dadurch aus, dass die Merkmalsvektoren in Traningsdaten vorklassifiziert sind, im Gegensatz zu **nicht überwachtem Lernen** (unsupervised learning) dessen Merkmalsvektoren nicht klassifiziert sind.

2.2 Wissensmodellierung

Die Wissensmodellierung hat das Ziel menschliches Wissen, dass zunächst nur implizit verfügbar ist zu formalisieren. Aus technischer Sicht kann die Wissensmodellierung als ein Prozess verstanden werden, der benötigtes Wissen zu einem gegebenen Problem in eine explizite, strukturierte Form bringt die den Bedürfnissen partizipierender Menschen und den Anforderungen der technologischen Umsetzbarkeit genügen (Kienreich und Strohmaier, 2006). Zu diesem Zweck existieren eine Reihen von Methoden die eine Organisation von Wissen ermöglichen und die sich in ihrer semantischen Reichhaltigkeit unterscheiden. Eine klassische Form der Wissensorganisation ist ein **kontrolliertes Vokabular**, bei dem ein Begriffssystem entsteht, dass darauf abzielt Homonyme[1] und

[1] Orthographisch gleiche Wörter mit unterschiedlicher Bedeutung (Mehrdeutigkeit) bspw. Bank (Sitzgelegenheit) - Bank (Finanz.)

Synonyme[2] zu vermeiden indem Konzepte genau einmal vorkommen oder zumindest auf einen bevorzugten Deskriptor abgebildet werden. Eine ausdrucksstärkere Methode bei der Wissensmodellierung ist der **Thesaurus**. Dieser besteht aus einer systematischen Sammlung von Konzepten zwischen denen thematische Beziehungen hergestellt werden. Ein Thesaurus enthält eindeutige Bezeichnungen für jedes Konzept wobei unterschiedliche Variatonen (Synonyme, Abkürzungen, Übersetzungen etc.) durch Äquivalenzrelationen miteinander in Beziehung gesetzt werden (Geyer-Hayden, 2009). Eine besonders ausdrucksstarke Form der Wissensmodellierung ist die **Ontologie**. Ursprünglich stammt der Begriff aus der Philosophie und beschreibt die Lehre des Seins. Die gängigste Definition aus informationstechnischer Sicht stammt von Gruber (Gruber, 1995), der eine Ontologie als

".. a formal, explicit specification of a shared conceptualization"

beschreibt. Hier bezieht sich Gruber auf den Begriff der Konzeptualisierung der eine abstrakte, vereinfachte Sicht auf den zu repräsentierenden Interessensbereich beschreibt und von einer Gruppe von Personen als geeignet betrachtet wird. Die Ausdruckstärke einer Ontologie variiert je nach Detailierungsgrad der Modellierung. Nach McGuinness (2002) reicht das Spektrum einer Ontologie von einem kontrollierten Vokabular bis hin zu komplexen Modellierungen mit Klassenzugehörigkeiten, disjunkten Klassen, Entität-Relationen oder Wert-Restriktionen.

2.2.1 Semantic Web

Das Web in seiner heutigen Form zeichnet sich durch eine unüberschaubar grosse Menge an Dokumenten aus die über Hyperlinks miteinander verbunden sind. Suchmaschinen nehmen eine Indizierung von Dokumenten vor, um diese durchsuchbar zu machen. Vordergründig ist das Web eine Entwicklung, die eine menschliche Interpretation der Daten vorsieht. Bei der Interpretation von Webseiten können menschliche Nutzer Inhalte meist problemlos erfassen, mit anderen Informationen in Beziehung setzen und in alternative Darstellungsformen transformieren. Die maschinelle Informationsverarbeitung hingegen ist nicht in der Lage dies ohne Weiteres zu bewerkstelligen, da die Semantik der Inhalte für Maschinen nicht erkennbar ist (Hitzler et al., 2008). Eine Möglichkeit Informationen im Web maschinell interpretierbar zu machen beschreibt das Konzept des *Semantic Web*, welches erstmals 2001 von Tim Berners-Lee und Anderen formuliert wurde (2001). Beschrieben wird das Semantic Web als Weiterentwicklung des heutigen

[2]Unterschiedliche Wörter mit gleicher Bedeutung Mobiltelefon - Handy

Web mit dem Ziel die maschinelle Interpretation von Daten auf Basis von wohldefinierten, offenenen Standards zu ermöglichen, um auf diese Weise die Zusammenarbeit zwischen Mensch und Maschine zu verbessern. Um dies zu erreichen muss Wissen, dass bisher nur in unstrukturierter Form existiert in formale, maschinell interpretierbare Form modelliert werden. Zu diesem Zweck existieren eine Reihe von Spezifikationen, die im Folgenden erläutert werden.

Ressource Description Framework

Das *Resource Description Framework*[3] (RDF) ist eine W3C[4] Spezifikation zur Repräsentation maschinenlesbarer Metadaten über Ressourcen im Web. Klassische Beispiele für Metadaten über Webressourcen sind Autor, Title oder Änderungsdatum. Durch eine Generalisierung des Konzepts von Webressourcen kann RDF zur Beschreibung beliebiger Ressourcen verwendet werden. Ressourcen werden über sogenannte *Unified Resource Identifier* (URI) identifiziert und referenziert. Ein einfaches Beispiel einer Modellierung findet sich in Abbildung 2.3, bei der auf die sogenannte *Friend of a Friend Ontologie*[5] (FOAF) zurückgegriffen wird. Die FOAF Ontologie erlaubt die Modellierung von Personen, ihren Beziehungen sowie ihre aktuellen Aktivitäten (bspw. Projekte). Für einen detaillierten Einblick über die FOAF Ontologie sei auf `http://xmlns.com/foaf/spec/` verwiesen. Die Abbildung zeigt die Modellierung von drei Personen die der FOAF Klasse *Person* angehören. Daneben wird eine weitere Resource mit Namen "SurgeryTube" modelliert, die der FOAF Klasse *Project* angehört. Über das FOAF Attribut *currentProject* lässt sich ausdrücken, dass die Personen Nils, Sebastian und Zeljko Projektmitglieder von SurgeryTube sind.

Die grundlegende Struktur in RDF ist das sogenannte *Triple*, welches sich aus einer Ressource, einem Prädikat und einem Objekt zusammensetzt. Auf diese Weise lassen sich einfache Aussagen über gegebene Sachverhalte ausdrücken. Zur Beschreibung eines RDF-Triple kommt die auf XML basierende Sprache RDF/XML zum Einsatz. Exemplarisch ist die Resource *Zeljko_ Carevic* aus Abbildung 2.3 in Listing 2.1 dargestellt. Die Beschreibung der Ressource "Zeljko_Carevic" erfolgt über das *rdf:about* Attribut. Als Triple ausgedrückt wird der Ressource "Zeljko_Carevic" über das Prädikat *foaf:name* das Objekt "Zeljko" zugewiesen. In diesem Fall ist das Objekt beschrieben durch eine einfache Zeichenfolge, kann jedoch auch Verweise auf weitere Ressourcen beinhalten.

[3] `http://www.w3.org/RDF/`
[4] World Wide Web Consortium
[5] `http://www.foaf-project.org/`

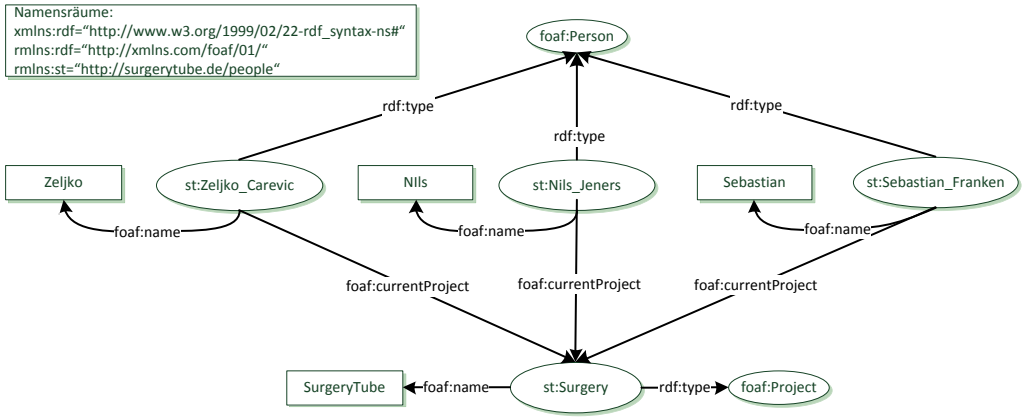


Abbildung 2.3: Beispiel einer Modellierung zu drei Personen und einem Projekt in FOAF

```
1 <?xml version="1.0"?>
2 <rdf:RDF
3 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
4 xmlns:foaf="http://xmlns.com/foaf/0.1/"
5 xmlns:st="http://SurgeryTube.de/People/">
6 <foaf:Person rdf:about="st:#Zeljko_Carevic">
7   <foaf:name>Zeljko</foaf:name>
8  </foaf:Person>
9 </rdf:RDF>
```

Listing 2.1: RDF-Modellierung mit FOAF

Linked Data

Zur Veröffentlichung und Vernetzung von wohldefinierten Daten auf Basis von Semantic Web Standards existiert eine Sammlung von Prinzipien, bekannt als Linked Data. Diese wurden 2006 erstmals von Berners-Lee formuliert (Berners-Lee, 2006) und lauten:

- Verwende URIs zur Benennung von Ressourcen
- Verwende HTTP URIs, damit menschiche Nutzer Ressourcen nachschlagen können
- Verwende Standards wie RDF oder SPARQL um nützliche Informationen zu präsentiert sobald jemand eine URI aufruft

- Binde Verweise auf andere URIs ein, damit auf diese Weise neues Wissen erlangt werden kann

Die zur Zeit wohl bekannteste Realisierung von Linked Data ist das *Linking Open Data* (LOD) Projekt[6] mit dem Ziel *open licence* Daten zu identifizieren und diese nach den Linked Data Prinzipien in RDF zu konvertieren und zu veröffentlichen. Die Adaption der Linked Data Prinzipien hat zu einem globalen Informationsraum geführt, in dem Daten aus verschiedenen Domänen wie Musik, Medizin, Firmen, Personen etc. verfügbar und maschinell interpretierbar gemacht wurden (Bizer et al., 2009). Neben der Konvertierung und Veröffentlichung von Daten nach dem Linked Data Prinzip steht das Erzeugen von semantischen Relationen zwischen Wissensbasen im Vordergrund. Angenommen man hat eine auf RDF basierte Wissensbasis über verschiedene Schauspieler (bspw. modelliert mittels FOAF), wohingegen eine weitere Wissensbasis Informationen über Hollywood Filme bereitstellt. Hier bietet es sich an, die beiden Wissensbasen mittels semantischen Relationen zu verknüpfen um auf diese Weise bereits modelliertes Wissen zu nutzen. Erst auf diese Weise entsteht ein Netz von semantisch aufbereiteten Informationen, die über Beziehungen miteinander verknüpft sind. Eine Form die verschiedenene Quellen, sowie ihre Relationen zueinander zu visualisieren ist die Linked Data Cloud, dargestellt in Abbildung 2.4. Inhalte in der Linked Data Cloud sind nach Domänen kategorisiert. Dazu zählen die Bereiche Medien, Geographie, Regierung, Publikationen, Lebenswissenschaften, User-generated Content und Cross-Domain. Insgesamt finden sich in der Linked Data Cloud 295 Datenquellen mit mehr als 31 Milliarden Tripplen (Bizer et al., 2011).

Als Schnittstelle zu Inhalten des Semantic Web bieten zahlreiche Wissensbasen einen sogenannten *SPARQL Endpoint*. Bei SPARQL[7] handelt es sich um eine Empfehlung des W3C[8]. SPARQL ist die Standard Abfragesprache des Semantic Web und erlaubt die gezielte Suche nach Inhalten in RDF Dokumenten.

2.3 PubMed

Bei PubMed[9] handelt es sich um eines der bekanntesten Werkzeuge bei der Literatursuche im Bereich der Medizin, welches zur Zeit von dem *National Center for Biotechnology Information* (NCBI) entwickelt und gewartet wird. PubMed umfasst mehr als 21 Millio-

[6] `http://linkeddata.org/`
[7] `http://www.w3.org/TR/rdf-sparql-query/`
[8] `http://www.w3.org/`
[9] `http://www.ncbi.nlm.nih.gov/pmc/`

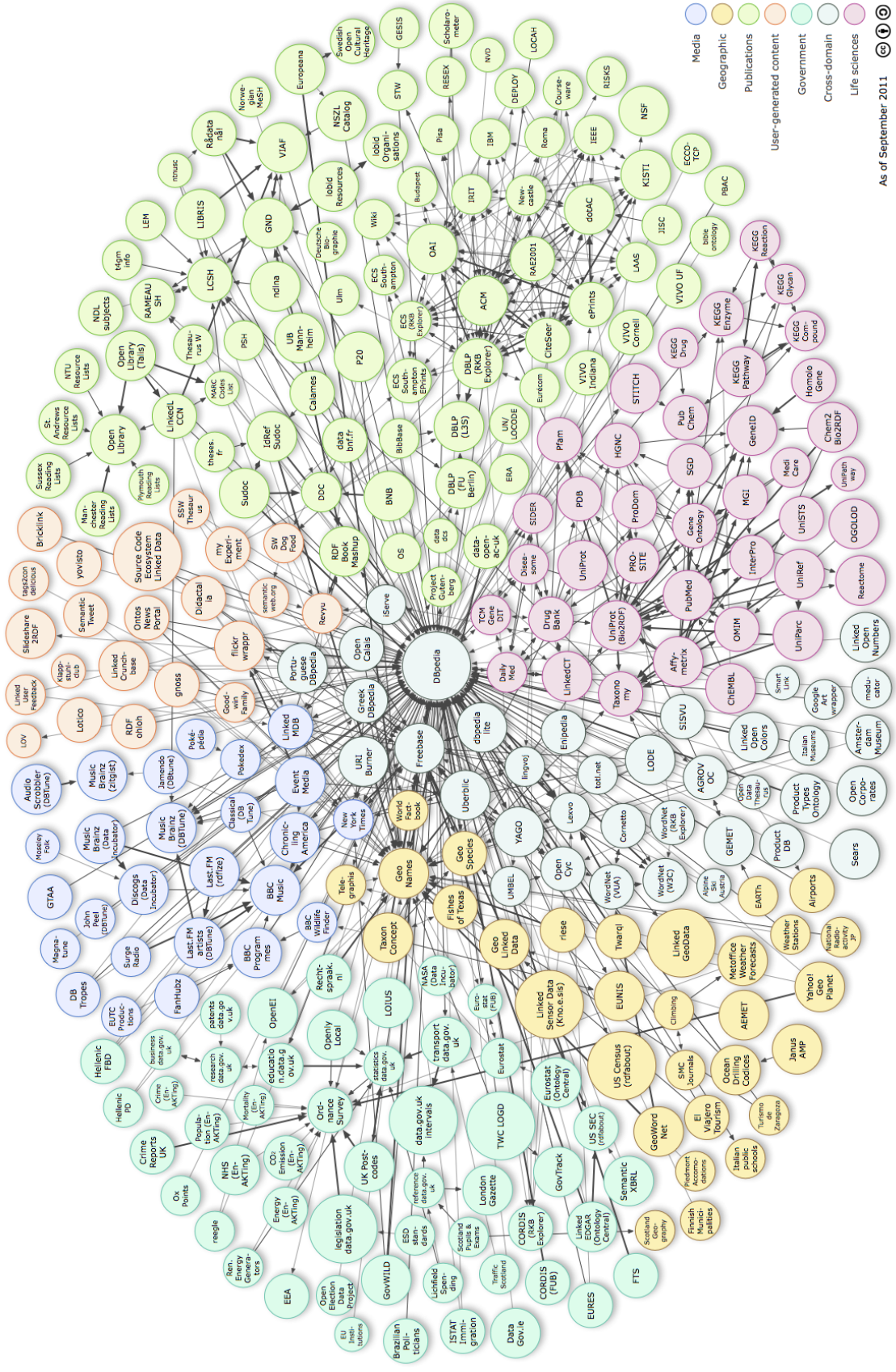


Abbildung 2.4: Linking Open Data Cloud Diagramm (Richard Cyganiak, 2011)

nen Aufsatzzitate aus biomedizinischer Literatur. Zu den Informationen die PubMed über die Aufsatzzitate veröffentlicht zählen eine kurze Zusammenfassung (Abstract) sowie Informationen über den Autor und die Form der Publikation. Es ist wichtig anzumerken, dass PubMed als bibliographischer Datenbestand vorgesehen ist der lediglich zusammenfassende Informationen über medizinische Publikationen bereitstellt. Wenn möglich finden sich jedoch Verweise auf Online-Quellen zu den Originalpublikationen. Bei der Indizierung der Inhalte verwendet PubMed ein kontrolliertes Vokabular genannt *Medical Subject Heading* (MeSH). Dies wird in Abschnitt 4.1 im Detail erläutert.

3 Stand der Technik

Der folgende Abschnitt soll einen Überblick über den Stand der Technik in den Bereichen der (semi-)automatischen Verschlagwortung und der Aggregation verteilter Wissensbasen zu einem zentralen Informationssystem bieten. Vorweg kann festgehalten werden, dass zwar zahlreiche Systeme identifiziert werden konnten, die sich mit der Verschlagwortung textbasierter Inhalte auseinandersetzten, jedoch werden diese nicht mit externen Wissensbasen in Verbindung gesetzt. Eine Konstellation wie sie in der vorliegenden Arbeit vorherrscht war so nicht zu finden. Entsprechend behandelt das folgenden Kapitel in Abschnitt 3.1 zunächst das automatische Verschlagworten von Inhalten. Davon losgelöst werden in Abschnitt 3.2 Systeme vorgestellt, die sich mit der Aggregation von unterschiedlichen semantisch aufbereiteten Wissensbasen beschäftigen. Zusammenfassend findet sich in 3.1.1 und 3.2.1 eine Übersicht über die betrachteten Systeme.

3.1 Verschlagwortung textbasierter Inhalte

Bei der Verschlagwortung textbasierter Inhalte konnten drei Vorgehensweisen identifiziert werden. Dazu zählt die Verschlagwortung mittels maschinellem Lernens, Empfehlungsdiensten und ein graphbasierter Ansatz. Im Folgenden werden die drei Ansätze, sowie entsprechende Referenzprojekte erläutert.

Verschlagwortung auf Basis maschinellem Lernens
Die Verschlagwortung textbasierter Inhalte ist ein viel diskutierter Bereich der in heutigen Systemen zumeist auf Basis von Algorithmen aus dem Bereich des überwachten Lernen realisiert wird. In Turney (1999) wird eine Extraktion sogenannter *Keyphrases* aus textbasierten Inhalten realisiert. Eine Keyphrase setzt sich aus einzelnen oder mehreren *Keywords*, im weiteren Verlauf als Schlagwörter bezeichnet, in einem Text zusammen. Turney betrachtet in seiner Arbeit die Verschlagwortung aus Sicht der Klassifikation, wobei jedes Wort oder jede Passage in einem Text entweder der Klasse Schlagwort oder nicht-Schlagwort angehört. In Witten und Andere (1999) wird ein ähnlicher Ansatz na-

mens *Keyphrase Extraction Algorithm* (KEA)[1] präsentiert, welcher von der Universität Waikato, Neuseeland entwickelt wurde. Bei der Klassifikation verwendet KEA einen zweidimensionalen Merkmalsvektor bestehend aus der *TF/IDF* Metrik[2] sowie der Distanz des Terms relativ zum Text. KEA unterscheidet zwischen der Extraktion von Schlagwörtern und Termzuweisung. Bei der Extraktion von Schlagwörtern wird davon ausgegangen, dass diese mindestens einmal innerhalb des analysierten Textes auftreten. Nach (Turney, 1999) ist dies in 70 bis 90 Prozent der Fall. Bei der Termzuweisung hingegen werden Schlagwörter innerhalb eines kontrollierten Vokabulars identifiziert und dem Dokument zugewiesen. Als Weiterentwicklung des KEA wird in Medelyan und Witten (2006) der KEA++ vorgestellt. KEA++ agiert als Mischung zwischen der Extraktion von Schlagwörtern und der Termzuweisung, genannt Schlagwort Indexierung. Hierbei werden zunächst Kandidaten aus einem Thesaurus identifiziert, welche einen Bezug zu dem Inhalt des Dokuments aufweisen. In einem zweiten Schritt erfolgt dann eine Filterung der zuvor ermittelten Kandidaten auf Basis maschinellen Lernens. Neben den bereits bekannten Merkmalen, TF/IDF und der Distanz des Terms relativ zum Text wurde KEA++ um zwei weitere Merkmale erweitert. Die Anzahl an Wörtern aus der sich der Schlagwort-Kandidat, zusammensetzt sowie einem durch semantische Relationen errechneten Knotenrang. Bei der Berechnung des Knotenrangs wird die Annahme verfolgt, dass ein Text der ein bestimmtes Thema behandelt auch mit hoher Wahrscheinlichkeit einen Großteil der Terme im Thesaurus zu diesem Thema umfasst. Es ist daher wahrscheinlich, dass sich signifikante Schlagwort-Kandidaten durch eine Vielzahl semantischer Relationen zu anderen Kandidaten auszeichnen. Eine Weiterentwicklung des KEA findet sich auch in Medelyan und Andere (2009) genannt Maui[3] wieder. Maui wurde wie auch KEA an der Universität Waikato entwickelt und basiert auf überwachtem Lernen. In einem ersten Schritt werden potentielle Kandidaten im Text identifiziert. Hier greifen verschiedene Regeln wie das Ignorieren von Termen, die lediglich einmal im Text auftreten oder auch Terme, die von einem *Stopword* umgeben sind. Insgesamt verwendet Maui neun Merkmale zur Extraktion von Schlagwörtern. Dazu zählt beispielsweise die Berechnung eines Knotenrangs. Ähnlich wie die Verwendung des Thesaurus in KEA++ wird in Maui ein semantische Relation errechnet wobei dies unter Verwendung von Wikipedia geschieht.

Verschlagwortung auf Basis von Empfehlungsdienst

Neben Ansätzen des maschinellen Lernens wird von Mishne in (2006) ein Ansatz zur Ex-

[1] `http://www.nzdl.org/Kea/index.html`

[2] Termfrequenz / Inverse Dokumentfrequenz beschreibt das Verhältnis der Worthäufigkeit in einem Dokument (TF) zu der Anzahl an Dokumenten die den Term enthalten (DF)

[3] `http://maui-indexer.googlecode.com`

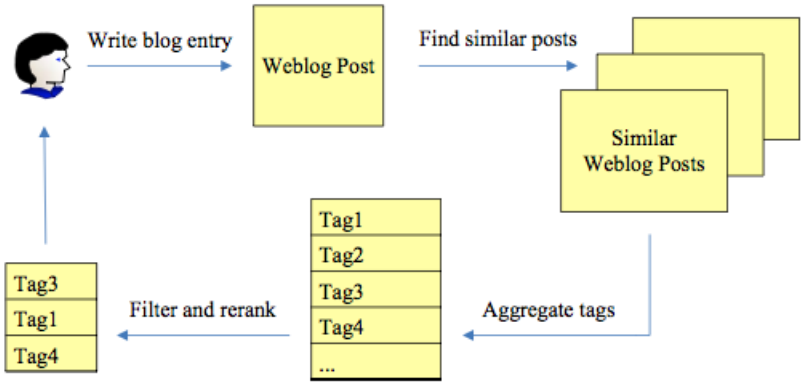


Abbildung 3.1: Automatische Verschlagwortung von Webblogs in AutoTag Mishne (2006)

traktion von Schlagwörtern auf Basis kollaborativer Filterungsmechanismen vorgestellt. Das sogenannte *AutoTag* verwendet Konzepte aus dem Bereich der Empfehlungsdienste (*Recommender Systems*) und präsentiert Nutzern potentielle Schlagworte, die bereits in ähnlichen Beiträgen verwendet wurden. Klassische Empfehlungsdienste verfolgen den Ansatz, dass ähnliche Nutzer auch ähnliche Interessen haben. Diese Annahme wird auch in AutoTag verfolgt, wobei hier die Perspektive verändert und davon ausgegangen wird, dass ähnliche Beiträge auch ähnlich verschlagwortet werden. Der Prozess der Verschlagwortung in AutoTag ist in Abbildung 3.1 dargestellt. Zur Identifikation ähnlicher Beiträge verwendet AutoTag Techniken des *Information Retrieval*. Nach der Analyse ähnlicher Beiträge findet eine Bewertung der verwendeten Schlagwörter in den N bestbewerteten Referenzblogs statt. Dies geschieht auf Basis der Häufigkeit mit der ein Term als Schlagwort in den Referenzblogs verwendet wurde. Abschliessend findet eine Neubewertung der Schlagwörter statt. Dies setzt vorraus das der Nutzer bereits eigene Schlagwörter vergeben hat, welche sich ebenfalls in der zuvor ermittelten Vorschlagsliste wiederfinden. Ist dies der Fall werden diese entsprechend stärker gewichtet.

Sood und Andere präsentieren in (2007) ein System namens *TagAssist*, das Nutzern Schlagworte bei der Erstellung von Webblogs vorschlägt, genannt *TagAssist* . Ähnlich wie AutoTag greift auch TagAssist auf Konzepte aus dem Bereich der Empfehlungsdienste zu. Bei der Erstellung eines neuen Webblogs werden ähnliche, bereits verschlagwortete Inhalte ermittelt. Die bereits vergebenen Schlagwörter werden aggregiert und bewertet. Die N bestbewerteten Schlagwörter werden dem Nutzer anschließend als mögliche Schlagwörter präsentiert. Im Gegensatz zu AutoTag ist der Nutzer in der Lage

auch selbst Schlagwörter zu etablieren die in die Vorschlagsliste aufgenommen werden.

Graphbasierte Ansätze
Zur Extraktion von Schlagwörtern stellen Mihalcea und Tarau in (2004) einen graphbasierten Ansatz, namens *TextRank* vor. Die Extraktion von Schlagwörtern basiert in TextRank auf dem aus Google bekannten *PageRank* Algorithmus (Brin und Page, 1998). Im Gegensatz zur Extraktion von Schlagwörtern auf Basis von maschinellem Lernen bzw. Empfehlungsdiensten erlaubt TextRank die Extraktion ohne Verwendung eines Dokumentkorpus respektive Referenzbeiträge. Die Extraktion erfolgt lediglich auf Basis des analysierten Dokuments. Die Annahme von Mihalcea und Tarau ist, dass ein Text in Form eines Graphen repräsentiert werden kann der dann unter Verwendung des PageRank Algorithmus bewertet wird. Nach einem optionalen Filterungsprozess werden Terme als Knoten in einem Graph repräsentiert. Die Ermittlung der Kanten zwischen zwei Knoten erfolgt auf Basis der Kookurenz zwischen Termen. Finden sich zwei Terme, die in einem gegebenen Fenster der Länge N auftreten werden diese im Graph durch eine Kante verbunden. Die Verwendung von Kookurenzen bei der Erzeugung des Graphen ist nicht vorgegeben, so dass jegliche Relationsformen zwischen Termen verwendet werden können.

3.1.1 Bewertung der Systeme

Bei Analyse des aktuellen Forschungsstands hat sich gezeigt, dass bereits verschiedene Ansätze existieren, die sich mit der automatischen Verschlagwortung von Inhalten beschäftigen. Die hier betrachteten Systeme sind zusammenfassend in Tabelle 3.1 dargestellt. Es lassen sich dabei drei verschiedene Vorgehensweisen identifizieren: Maschinelles Lernen, Techniken aus dem Bereich der Empfehlungsdienste und graphbasierte Ansätze. Die Extraktion von Schlagwörtern durch maschinelles Lernen setzt große Mengen an Trainingsdaten vorraus. Bei Nutzung der Empfehlungsdienste wird eine Plattform vorrausgesetzt, die ausreichend Referenzbeiträge zur Verfügung stellt. Einzig der graphbasierte Ansatz in TextRank kommt ohne jegliche Abhängigkeiten aus.

3.2 Aggregation medizinischer Inhalte

Bei der Aggregation medizinischer Inhalte können drei Systeme identifiziert werden, die sich in ihrem Umfang und den verwendeten Wissensbasen unterscheiden. Diese werden

System	Algorithmus	# Merkmale	Kontrolliertes Vokabular
AutoTag	Empfehlungsdienste	2	Nein
KEA	Maschinelles Lernen	2	Ja
KEA++	Maschinelles Lernen	3	Ja
Maui	Maschinelles Lernen	9	Ja
TagAssist	Empfehlungsdienste	2	Nein
TextRank	PageRank	-	Ja

Tabelle 3.1: Übersicht der betrachteten Systeme zur automatischen Verschlagwortung

im Folgenden kurz erläutert und in Abschnitt 3.2.1 abschliessend miteinander verglichen und bewertet.

Linking Open Drug Data

Innerhalb des sogenannten *Linking Open Drug Data*[4] (LODD) Projekts wurden Datenquellen zu Medikamenten, klinischen Studien, Krankheiten, traditioneller chinesische Medizin und pharmazeutischen Unternehmen nach den Linked Data Prinzipien konvertiert und in die Linked Data Cloud integriert. LODD ist daher als Teilmenge der Linked Data Cloud zu verstehen. Das Ergebnis der Konvertierung besteht aus insgesamt 8 Millionen Triple und knapp 400.000 Verknüpfungen zu externen Daten, die zumeist über sogenannte *sameAs* Relationen realisiert sind (Jentzsch et al., 2009). In Abschnitt 6.2.2 wird das LODD Projekt im Detail erläutert.

GoPubMed

Bei GoPubMed[5] handelt es sich um ein System zur Verbesserung der Literatursuche in PubMed. GoPubMed erlaubt dem Nutzer eine ontologiebasierte Suche nach Pubmed-Artikeln und präsentiert kategorisierte Ergebnisse. Die Kategorisierung erfolgt in 4 Top-Level Kategorien: Was, Wer, Wo und Wann. In Abbildung 3.2 ist eine beispielhafte Suche nach dem Begriff *Alzheimer Disease* dargestellt. Neben der Kategorisierung und den Pubmed Artikeln stellt GoPubMed zudem einige Statistiken zu dem gesuchten Begriff dar sowie eine kurze Beschreibung. Als Grundlage zur Kategorisierung verwendet GoPubMed das kontrollierte Vokabular MeSH, sowie eine Ontologie über Genprodukte[6]. GoPubMed basiert auf der Dissertation von Andreas Dorm (2009) und wird zur Zeit von Transinsight verwaltet.

[4] `http://www.w3.org/wiki/HCLSIG/LODD`
[5] http://www.Gopubmed.org
[6] `http://www.geneontology.org/`

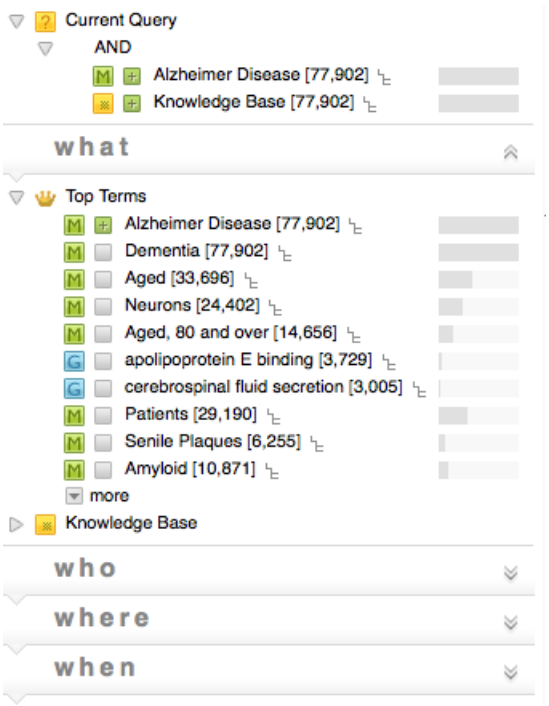


Abbildung 3.2: GoPubMed - Suche nach dem Konzept Alzheimer Disease. Quelle: (Transinsight, 2012)

Pathway and Interaction Knowledge Base

Bei *Pathway and Interaction Knowledge Base* (PIKB) handelt es sich um eine Ontologie, die zahlreiche Quellen aus dem Bereich der Bioinformatik und Biomedizin semantisch integriert. Dazu zählen unter anderem PubMed, sowie Inhalte aus dem LODD Projekt (Momtchev et al., 2009).

Als prototypische Implementation von PIKB wurde Linked Life Data[7] entwickelt, welches sich zur Zeit[8] noch in einer Betaphase befindet. Linked Life Data aggregiert Informationen zu einem Konzept aus verschiedenen Quellen und stellt diese dem Nutzer dar. Zu den Informationen zählen beispielsweise Synonyme, eine kurze Beschreibung des gesuchten Konzepts oder auch Inhalte aus PubMed. Bei der Bildung von Relationen zwischen Konzepten kommen Informationen aus *DBPedia*[9] zum Einsatz. Hier werden *WikiLink* Verbindungen zwischen Themen zur Bildung von Ressource-Relationen verwendet. Zusätzlich kommen Algorithmen zum Einsatz, die Gemeinsamkeiten zwischen Ressourcen ermitteln indem eine Inhaltsanalyse durchgeführt wird. Die Web-Oberfläche bietet neben einer klassischen Navigation durch Inhalte und einer Suche auch einen

[7] http://linkedlifedata.com/
[8] Stand 04/2012
[9] `http://dbpedia.org/About` Eine in RDF konvertierte Version von Wikipedia

SPARQL Endpoint an, der externe Anfragen an das System erlaubt. In der aktuellen Version 0.9 von Linked Life Data sind mehr als 20 verschiedene Quellen mit insgesamt mehr als 1 Milliarde[10] Ressourcen integriert.

3.2.1 Bewertung der Systeme

Die in 3.2 dargestellte Tabelle soll einen Überblick über die behandelten Systeme im Bereich der Aggregation semantischer Inhalte liefern. Nach Betrachtung der verschiedenen Systeme wird deutlich das zwischen PIKB und der hier vorliegenden Zielsetzung Gemeinsamkeiten vorliegen. PIKB aggregiert Informationen aus verschiedenen Quellen in einem System und stellt diese dem Nutzer dar. Neben Information über das Konzept, seine Relationen und Synonyme aggregiert PIKB Informationen aus PubMed.

System	# Quellen	# Konzepte	Relationen	Lizenz
GoPubMed	2	circa 130.000	-	proprietär
LODD	6	?	388.000	Open Licence
PIKB	23	circa 1 Mrd.	circa 15 Mrd.	Open Licence

Tabelle 3.2: Übersicht verschiedener System zur Aggregration medizinischer Inhalte

[10]Stand 03/2012 `http://linkedlifedata.com/sources`

4 Eignung medizinischer Begriffssysteme

Wissen in der Medizin ist im Wesentlichen in sprachlicher Form dokumentiert. Die verwendete Fachterminologie ist dabei besonders in Spezialbereichen inkonsistent was dazu führte, dass Standardisierungsbemühungen unternommen wurden, welche medizinisches Wissen in rechnergestützen Systemen dokumentieren. Im Fokus steht dabei die Etablierung eines kontrollierten Vokabulars und die einhergehende Strukturierung in Begriffsordnungen (Spreckelsen und Spitzer, 2009). Ein durch Experten erzeugtes kontrolliertes Vokabular kann für Mediziner von enormem Nutzen sein, da es die Möglichkeit schafft in einer unüberschaubar großen Menge an Daten genau die zu identifizieren, die benötigt werden. Ist einem Nutzer von PubMed beispielsweise das Ordnungssystem und die Terminologie bekannt, so ist er in der Lage gezielt Inhalte aufzufinden und erhöht somit die Präzision seiner Suchergebnisse. Dennoch existiert auch Kritik an der Nutzung eines kontrollierten Vokabulars als Grundlage für PubMed. So verspricht die Nutzung eines kontrollierten Vokabulars, verglichen mit einer Schlüsselwort-Suche einen verbesserten Zugriff auf Inhalte, da diese konsistent indiziert sind. Jedoch fällt Nutzern das Mapping des Suchbegriffs auf die verwendete Terminologie schwer (Gault et al., 2002).

Ein Aspekt mit dem sich die vorliegende Arbeit auseinandersetzt ist, wie mit Hilfe medizinischer Begriffssysteme textbasierte Inhalte klassifiziert werden. Im Folgenden Kapitel werden zwei medizinische Begriffssysteme untersucht die als zugrundeliegendes Vokabular infrage kommen. Zwei grundsätzliche Anforderungen lassen sich dabei vorab festhalten. Zum einen muss sichergestellt werden, dass eine Zuordnung der Begriffe über Sprachgrenzen hinweg möglich ist, da externe Inhalte wie PubMed zumeist in englischer Sprache verfasst sind. Zum anderen muss das Begriffssystem möglichst umfassend sein um nicht nur grundlegende, sondern auch tiefergehende Begriffe identifizieren zu können. Ein solches Begriffssystem könnte ein Thesaurus sein der neben der reinen Begriffssammlung zusätzlich das Abbilden von Beziehungen zwischen Konzepten erlaubt. Es werden nun die Begriffssystem *Medical Subject Heading* (MeSH) und das

Unified Medical Language System (UMLS) im Detail erläutert und in Abschnitt 4.3 auf ihre Einsatzmöglichkeiten hin verglichen.

4.1 Medical Subject Heading

Das *Medical Subject Heading* (MeSH)[1] ist ein 1960 von der National Library of Medicine (NLM) erstelltes kontrolliertes Vokabular welches jährlich aktualisiert wird. In der englischsprachigen Version von 2011 umfasst MeSH mehr als 26.000 Hauptbegriffe und über 180.000 Synonyme (MeSH Fact Sheet, 2012). Die NLM verwendet MeSH primär zur Indexierung von PubMed Artikeln. Dies macht MeSH zu einem interessanten Hilfsmittel bei der Suche nach Inhalten in PubMed. So wird es möglich gezielt nach einem MeSH Konzept zu suchen oder auch komplexere Suchanfragen zu erstellen bei denen auch verwandte Begriffe bzw. Synonyme berücksichtigt werden. Aus diesem Grund eignet sich MeSH besonders dann, wenn ein konsistenter Abgleich zwischen textbasierten Inhalten und PubMed geplant ist. Wird ein Blogeintrag bspw. mit dem Begriff "Alzheimer-Krankheit" verschlagwortet kann dies auf das Mesh Konzept *Alzheimer-Disease* abgebildet werden womit eine konsistente Integration von PubMed Inhalten gewährleistet werden kann.
MeSH organisiert Begriffe hierarchisch in der sogenannten *MeSH Tree Structur* mit insgesamt 16 Hauptkategorien. Zu den Hauptkategorien zählen beispielsweise Anatomie, Organismus, Krankheiten, Informationswissenschaft oder Geographie. Betrachtet man die Hauptkategorie Geographie welche Kontinente, Länder und Städte beinhaltet wird deutlich, dass sich in MeSH auch Konzepte finden die sich nicht ausschliesslich dem Bereich der Medizin zuordnen lassen. Zu jedem Konzept innerhalb des MeSH existieren verschiedene Attribute welche dieses näher beschreiben. Besonders interessant für diese Arbeit sind der Hauptbegriff, eine kurze Beschreibung, Synonyme sowie die *Tree Number* eines jeden Konzepts. Bei der Tree Number handelt es sich um durch Punkte getrennte Knoten innerhalb der MeSH Tree Structure. Für das MeSH Konzept *Alzheimer-Disease* lautet die MeSH Tree Structur beispielsweise:

- Nervous System Diseases [C10] - Central Nervous System Diseases [C10.228] - Brain Diseases [C10.228.140] - Dementia [C10.228.140.380] - Alzheimer Disease [C10.228.140.380.100]
- Nervous System Diseases [C10] - Neurodegenerative Diseases [C10.574] - Tauopathies [C10.574.945] - Alzheimer Disease [C10.574.945.249]

[1] http://www.nlm.nih.gov/mesh/meshhome.html

- Mental Disorders [F03] - Delirium, Dementia, Amnestic, Cognitive Disorders [F03.087] - Dementia [F03.087.400] - Alzheimer Disease [F03.087.400.100]

Ein Ausschnitt aus dem MeSH Konzept *Alzheimer-Disease* ist in Tabelle 4.1 dargestellt (der vollständige Eintrag findet sich im Anhang 8.1). Es wird deutlich, dass der Term *Alzheimer-Disease*, wie bereits dargestellt innerhalb von 3 Kategorien zu finden ist. Weiterhin ist zu erkennen, dass neben dem Hautbegriff (Main Heading) auch Synonyme (Entry Term) abgebildet werden, die unterschiedliche Schreibweisen berücksichtigen und das zu jedem Term eine kurze Beschreibung existiert (Scope Note).

Attribut	Wert
Main Heading	Alzheimer Disease
Scope Note	A degenerative disease of the BRAIN characterized by the insidious onset of DEMENTIA. Impairment of MEMORY, judgment, attention span, and problem solving skills are followed by severe APRAXIAS and a global loss of cognitive abilities. The condition primarily occurs after age 60, and is marked pathologically by severe cortical atrophy and the triad of SENILE PLAQUES; NEUROFIBRILLARY TANGLES; and NEUROPIL THREADS. (From Adams et al., Principles of Neurology, 6th ed, pp1049-57)
Tree Number	C10.228.140.380.100
Tree Number	C10.574.945.249
Tree Number	F03.087.400.100
Entry Term	Alzheimer's Disease
Entry Term	Alzheimer's Disease, Focal Onset
Entry Term	Dementia, Alzheimer Type
Entry Term	Dementia, Senile

Tabelle 4.1: Verkürzter Beispieleintrag MeSH 2011 Alzheimer Krankheit

Neben der standardmäßig in Englisch verfassten Version des MeSH existiert eine deutsche Übersetzung. Für die Übersetzung in die deutsche Sprache ist das DIMDI[2] (Deutsches Institut für Medizinische Dokumentation und Information) verantwortlich. Bei der Übersetzung werden neben den Hauptbegriffen auch Synonyme berücksichtigt. Eine Übersicht der enthaltenen Deskriptoren ist in Abbildung 4.1 dargestellt. Es ist zu erkennen, dass die Anzahl an Hauptschlagwörtern in beiden Sprachen identisch ist und somit zu jedem englischen Hauptschlagwort eine entsprechende deutschsprachige Übersetzung exisitiert. Bei der Übersetzung von englischsprachigen Synonymen findet sich jedoch ein deutlicher Unterschied. Hier wurden nur knapp ein Drittel der über 180.000

[2] `http://www.dimdi.de/static/de/klassi/mesh_umls/mesh/index.htm`

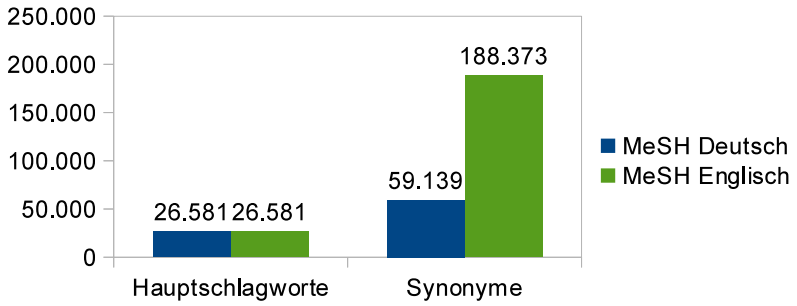


Abbildung 4.1: Deskriptoren in der deutschen und englischen Version des MeSH nach (DIMDI, 2012)

englischsprachigen Synoyme in die deutsche Sprache übersetzt. Weiterhin bleiben Inhalte wie die Beschreibung eines Konzepts in der Originalfassung erhalten.

4.2 Unified Medical Language System

Das *Unified Medical Language System*[3] ist ein seit 15 Jahren entwickeltes Verzeichnis biomedizinscher Begriffe welches von der *National Library of Medicine* (NLM) entwickelt wurde. Das Ziel von UMLS ist die Schaffung eines einheitlichen Vokabulars um die Menge an Begriffen zu verringern die für ein Konzept existieren und um ein einheitliches Datenformat zu definieren (Bodenreider, 2004). Über die Zeit haben sich in der Medizin zahlreiche Begriffssammlungen entwickelt, die sich über verschiedene Sprachen und medizinische Teilgebiete erstrecken. Zwangsläufig ergeben sich dabei auch redundante Bereiche mit teils unterschiedlichen Begrifflichkeiten. Diese werden in UMLS zu einem gemeinsamen Konzept kombiniert. Betrachtet man die Krankheit Alzheimer, welche sich bspw. innerhalb von MeSH unter dem Konzept *Alzheimer-Disease* und innerhalb des ICD[4] (International Classification of Diseases) unter *Alzheimer's disease* findet, wird deutlich, dass unterschiedliche Quellen thematische Überschneidungen besitzen. UMLS kombiniert solche Überschneidungen unter einem generalisierten Konzept, welches zur Zeit[5] mehr als 100 biomedizinische Quellen mit mehr als 10 Mio. Begriffe für knapp 2,5 Mio. Konzepte umfasst (UMLS Statistics NLM, 2012). Darunter findet sich auch eine mehrsprachige Fassung des MeSH. Die verschiedenen Quellen sind in Abbildung 4.2 vereinfacht dargestellt.

[3] http://www.nlm.nih.gov/research/umls/
[4] `http://www.who.int/classifications/icd/en/`
[5] (Stand 06/2011)

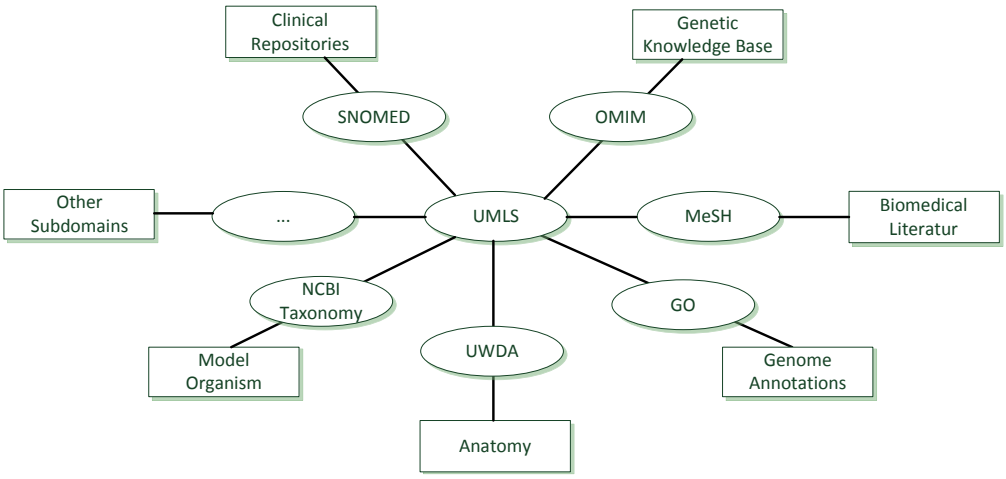


Abbildung 4.2: Vereinfachte Darstellung der UMLS Quellen (Bodenreider, 2004)

UMLS wird von der National Library of Medicine gepflegt und ist nach einer Registrierung als Download oder auch als Webservice verfügbar. Das UMLS setzt sich aus drei sogenannten *Knowledge Tools* zusammen (UMLS Overview, 2008):

- Metathesaurus
  Der Metathesaurus ist das Kernstück des UMLS und umfasst die Begriffssamlung aus den o.g. Quellen und die Beziehungen zwischen Konzepten

- Semantic Network
  Innerhalb des Semantic Network finden sich semantische Informationen zu verschiedenen Klassen und ihren Beziehungen. Eine Klasse könnte bspw. eine Krankheit sein wobei über semantische Relationen das Verhältnis zwischen Klassen ausgedrückt wird.

- SPECIALIST Lexicon und Lexical Tools
  Das SPECIALIST Lexicon ist ein englischsprachiges Lexikon aus dem Bereich der Biomedizin. Bei den Lexical Tools handelt es sich um eine Sammlung zahlreicher Werkzeuge zur Verarbeitung von natürlicher Sprache

4.3 Bewertung

Beide hier vorgestellten Begriffssysteme eignen sich für den Einsatz in Anwendungen zur Klassifikation textbasierter Inhalte. Je nach Umfang und Zielsetzung kann entweder das umfangreichere UMLS oder das MeSH verwendet werden. MeSH ist im Vergleich zu UMLS deutlich schlanker und eignet sich besonders dann wenn eine Integration von PubMed Inhalten geplant ist. Im Vergleich zu MeSH ist UMLS wesentlich umfangreicher und aggregiert zahlreiche verschiedene medizinische Quellen (darunter auch MeSH) in unterschiedlichen Sprachen zu einem einheitlichen Konzept. Dadurch sind Systeme die UMLS einsetzen in der Lage eine Vielzahl von unterschiedlichen Schreibweisen zu erkennen und dann entsprechend einem Konzept zuzuordnen. Durch seinen Umfang stellt UMLS einen erheblich größeren Aufwand bei der Integration und Wartung dar, jedoch besitzt es Potential für eine zusätzliche Systemunterstützung die über die Klassifikation von Inhalten hinausgeht. Betrachtet man das Potential des UMLS wird deutlich, dass weitere Informationen für zukünftige Nutzer von Wert sein können. Denkbar wäre hier beispielsweise ein intensiver Umgang mit Relationen oder auch das Integrieren von englischsprachigen Quellen. Hier bietet sich Raum für Folgearbeiten die sich intensiver mit dem UMLS und seinen verschiedenen Quellen auseinandersetzen. Da MeSH als Quelle in UMLS integriert ist und der Umfang von UMLS zudem Potential für weitere Systemunterstützungen bietet wird für den weiteren Verlauf UMLS speziell der UMLS Metathesaurus als Begriffssystem verwendet. Durch die Nutzung von UMLS können, im Vergleich zu einer Beschränkung auf MeSH deutlich mehr unterschiedliche Schreibweisen zu einem Konzept identifiziert werden, was sich als hilfreich bei der automatischen Verschlagwortung von Inhalten erweisen kann. Weiterhin garantiert es durch die Nutzung einer einheitlichen Terminologie eine präzise Integration von PubMed Inhalten.

4.4 Aufbau des UMLS Metathesaurus

Der UMLS Metathesaurus unterscheidet zwei verschiedene relationale Formate, dem *Rich Release Format*(RRF) sowie dem *Original Release Format*(ORF). Dabei wird für Entwickler der Einsatz des RRF empfohlen welches signifikante Vorteile in der Transparenz der Quellvokabularien besitzt (UMLS Reference Manual NLM, 2011). Für ein besseres Verständnis der Folgeinhalte ist es an dieser Stelle notwendig den genauen Aufbau des UMLS Metathesaurus zu erläutern. Besonders hervorgehoben wird hier-

bei die Organisation von Konzepten innerhalb des UMLS Metathesaurus und welche Datenbanken für eine Identifikation relevant sind.

Ein Ziel des UMLS Metathesaurus ist das Verbinden der verschiedenen Quellen und ihrer Begriffe zu einem einheitlichen Konzept unter Abbildung der entsprechenden Synonyme. Zu diesem Zweck ist der UMLS Metathesaurus nach Konzepten organisiert, wobei ein Konzept im Kontext des UMLS als die Menge aller Bezeichner verstanden werden kann. Sämtliche Konzepte werden in der Datenbank *MRCONSO* verwaltet und über einen eindeutigen *CUI* (Concept Unique Identifier) identifiziert. Der CUI agiert dabei als zentrales Element mit dessen Hilfe Relationen, Bezeichnungen und Attribute verknüpft werden können. Neben dem CUI finden sich zahlreiche weitere Informationen innerhalb der Datenbank, wie die Quelle des Konzepts oder auch die Sprache der Konzeptbezeichnung. Hierbei ist anzumerken, dass ein CUI zwar ein Konzept eindeutig identifiziert jedoch besteht zwischen Konzepten und ihren Quellen, Sprachen und Bezeichnungen eine 1:n Beziehung. So ist das Konzept *Alzheimer Disease* im Metathesaurus über den CUI "C0002395" identifiziert. Das Konzept existiert in mehreren Quellen, in unterschiedlichen Schreibweisen und unterschiedlichen Sprachen jedoch verweisen alle auf das generelle Konzept CUI "C0002395". Dies ist in Tabelle 4.2 vereinfacht dargestellt und auf die zwei Quellen MeSH (Deutsch) und MEDLINEPLUS beschränkt[6]. Es ist zu erkennen, dass im deutschsprachigen MeSH insgesamt sieben verschiedene Schreibweisen zu dem Konzept *Alzheimer Disease* existieren und in MEDLINEPLUS zwei Schreibweisen. Beide Quellen verweisen jedoch auf denselben CUI.
Zusätzlich zu der Abbildung von Synonymen erlaubt der Metathesaurus auch die Verknüpfung von Beziehungen zwischen Konzepten. Beziehungen werden dabei von der Datenbank *MRREL* verwaltet. Dabei wird zwischen Intra- und Inter-Quellbeziehungen unterschieden. Bei der Bildung von Intra-Quellbeziehungen werden Relationen primär aus dem Quellvokabular hergestellt. Dies geschieht entweder explizit über das Quellvokabular oder implizit beispielsweise durch hierarchische Nähe. Dazu existieren Intra-Quellbeziehungen die durch statistische Berechnung gewonnen werden indem analysiert wird wie häufig Konzepte gemeinsam als Schlagwörter auftreten. Beispielsweise lassen sich Intra-Quellbeziehungen in MeSH identifizieren indem betrachtet wird wie häufig Konzepte gemeinsam als Schlagwörter in PubMed verwendet werden. Solche Arten der Relation werden in der Datenbank *MRCOC* verwaltet (UMLS Reference Manual NLM, 2011).

[6]Das vollständige Konzept *Alzheimer Disease* findet sich im Anhang 8.1

CUI	SAB	STR
C0002395	MSHGER	Alzheimer-Krankheit
C0002395	MSHGER	Dementia senilis
C0002395	MSHGER	Demenz, senile
C0002395	MSHGER	Demenz, Alzheimer-Typ
C0002395	MSHGER	Senile Demenz, Alzheimer-Typ
C0002395	MSHGER	Demenz, primär degenerative senile
C0002395	MSHGER	Demenz, Primär degenerative senile Demenz
C0002395	MEDLINEPLUS	Alzheimer's Disease
C0002395	MEDLINEPLUS	AD

Tabelle 4.2: Verkürzter Beispieleintrag zu dem UMLS Konzept Alzheimer Disease

4.5 UMLS Metathesaurus Vorverarbeitung

Bei der Verwendung des UMLS Metathesaurus wurden einige Vorverarbeitungsprozesse durchgeführt. Die Installation des UMLS erlaubt eine skriptbasierte Extraktion der Inhalte in verschiedene Datenbanksysteme[7]. Für den weiteren Verlauf der Arbeit wird ein Export nach MySQL[8] gewählt. Nachdem das Skript erfolgreich ausgeführt wurde liegen die ausgewählten Inhalte als MySQL Tabellen vor. Um den Berechnungs- und Wartungsaufwand möglichst gering zu halten wurde eine Filterung auf bestimmte Inhalte vorgenommen. Im Speziellen liegt der Fokus auf der deutschsprachigen Version des MeSH. Grundsätzlich von Interesse sind neben den deutschsprachigen Bezeichnungen von Konzepten, ihre Hierarchie sowie ihre Relation zu anderen Konzepten. Bei der Ermittlung medizinisch relevanter Begriffe in einem Text stellt sich die Frage wie ein Abgleich zwischen Inhalten und dem UMLS Metathesaurus erfolgen kann. Hier wird zwischen zwei Übereinstimmungskriterien unterschieden:

- Enthaltene Zeichenfolge
  Es werden alle Begriffe als Ergebnis zurückgegeben welche die zu suchende Zeichenkette am Anfang oder Ende einer Konzeptbezeichnung beinhalten.

- Exakte Übereinstimmung der Begriffe
  Es werden nur Konzepte ermittelt die mit dem zu vergleichenden Token vollständig übereinstimmen.

Sucht man innerhalb des Metathesaurus nach dem Konzept "Fettleber" erhält man für die zwei Übereinstimmungskriterien die in Tabelle 4.3 dargestellten Ergebnisse. Eine

[7] Eine ausführliche Beschreibung der verschiedenen Formate findet sich in `http://www.nlm.nih.gov/research/umls/implementation_resources/scripts/index.html`

[8] `http://www.mysql.de/`

	Enthaltene Zeichenfolge	**Exakte Übereinstimmung**
CUI	Konzeptbezeichnung	Konzeptbezeichnung
C0015695	Fettleber	Fettleber
C0015695	Fettleber,alkoholisch	-

Tabelle 4.3: Ergebnisse der Suche nach *Fettleber* für die Übereinstimmungskriterien

Suche nach enthaltener Zeichenfolge würde die Konzeptbezeichnungen "Fettleber" und "Fettleber,alkoholisch" ermitteln welche beide auf das generelle Konzept "C0015695" verweisen. Eine Suche nach exakter Übereinstimmung liefert lediglich die Konzeptbezeichnung "Fettleber" als Ergebnis.

Beide Varianten der Übereinstimmungskriterien bieten Vor- und Nachteile. Bei einer strengen Übereinstimmung kann davon ausgegangen werden, dass eine vergleichbar geringe Menge an passenden Konzepten gefunden wird. Dies birgt jedoch die Gefahr, dass Übereinstimmungen aufgrund von geringen Abweichungen in der Schreibweise nicht identifiziert werden können. Verwendet man die weniger strengen Übereinstimmungskriterien erhält man analog dazu eine große Menge an Konzepten. Dies bietet ein hohes Maß an Vollständigkeit, führt jedoch gleichzeitig zu einem Qualitätsverlust in der Präzision mit der Konsequenz, dass große Mengen an irrelevanten Begriffen in die Weiterverarbeitung gelangen können.

Ein Vorteil der zweiten Variante wird jedoch erst auf den zweiten Blick ersichtlich. Mit einem weniger strengen Übereinstimmungskriterium erhält man die Möglichkeit auf Begriffe zu stoßen die sich nur indirekt aus dem Text ergeben. Dies lässt sich an einem Beispieltext zum Thema "Fettleber" verdeutlichen welcher im Anhang 8.2 dargestellt ist. Der Text wurde mit den Begriffen "Alkoholismus", "Alkoholsucht" und "Leber" verschlagwortet. Dabei werden die Begriffe "Alkoholismus" und "Alkoholsucht" im Text nicht erwähnt. Dies macht es für die strenge Variante der Übereinstimmung unmöglich diese Begriffe zu identifizieren. Da der Begriff "Alkohol" jedoch in dem Text vorkommt würde die zweite Variante die Zeichenfolge Alkohol in dem gültigen Konzept *C0001973 Alkoholismus* identifizieren und als Teil des Ergebnisses zurückliefern. Aus diesem Grund wird für die weitere Verarbeitung trotz der großen Menge an Konzepten die Variante mit dem weniger strengen Übereinstimmungskriterium gewählt.
Eine Auffälligkeit des UMLS ist die Menge an Inhalten die nur bedingt medizinisch relevant sind. So finden sich im UMLS zahlreiche Inhalte die sich nur indirekt auf die Medizin zurückführen lassen. Dabei sind Geographische Konzepte, Tätigkeiten uvm. zu nennen. Eine Filterung der verwendeten Inhalte im UMLS kann an dieser Stelle jedoch

nicht vorgenommen werden da die Relevanz eines Konzepts nur durch Domänenexperten beurteilt werden kann.

5 Automatische Verschlagwortung textbasierter Inhalte

In der letzten Jahren vollzog sich im Web ein Paradigmenwechsel der den früheren Kosumenten von Inhalten zugleich zum Produzenten machte (dem sogenannten *Prosumer*), was zu einem erheblichen Anstieg an nutzergenerierten Inhalten (User generated content) führte (O'Reilly, 2005). Gleichzeitig wächst auch der Bedarf an einer Organisation der neu erschlossenen Inhalte. Eine Möglichkeit Inhalte zu organisieren ist das Verschlagworten (Tagging)[1], bei dem Inhalte mit Metadaten angereichert werden. Inhalten werden eine Reihe von Schlagworten (Tags) zugewiesen mit dessen Hilfe sich das behandelte Thema beschreiben lässt. Auf diese Weise bildet sich eine Begriffssammlung, die eine Suche, Organisation und Navigation durch Inhalte erlaubt (Sood et al., 2007). In Abbildung 5.1 ist eine *Tag Cloud* dargestellt, die zeigt wie innerhalb einer Musikplattform ein bestimmter Künstler verschlagwortet ist. Auf diese Weise können Nutzer anhand von Schlagwörtern navigieren, um zu ähnlichen Inhalten zu gelangen und wie in diesem Fall beispielsweise auf neue Künstler zu stoßen. Zu den Systemen die eine Verschlagwortung unterstützen zählen unter anderem *Flickr*[2] (Verschlagwortung von Bildern), *YouTube*[3] (Verschlagwortung von Videos) oder *Blogspot*[4] (Verschlagwortung von Blog Einträgen).

Traditionelle Kategoriesierungen, wie man sie beispielsweise in Bibliotheken findet, werden in der Regel durch eine Reihe von Experten vorgenommen. Im Gegensatz dazu verfolgt die Verschlagwortung das Prinzip, dass jeder Inhalte frei annotieren kann (Golder und Huberman, 2006). Auf diese Weise ensteht ein informelles System an Begriffen, dass sich ohne grossen Aufwand erstellen lässt. Im Jahr 2011 zeigte sich laut Technorati[5], dass über 60% aller Blog-Autoren Schlagworte zur Beschreibung ihrer Einträge nutzen. Unter den Top 100 Bloggern sind es 92% die Schlagworte verwenden.

[1] Auch bezeichnet als *collaborative tagging*, *social classification*, *social indexing*, *folksonomy*(Voss, 2007)
[2] `http://www.flickr.com/`
[3] `http://www.youtube.com/`
[4] `http://googleblog.blogspot.com/`
[5] `http://technorati.com/`

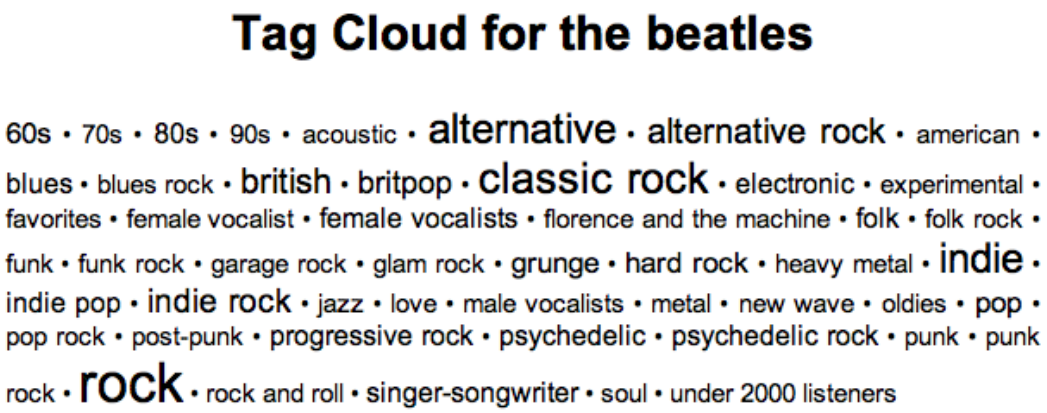


Abbildung 5.1: Tag Cloud zu den Beatles. Quelle: `www.lastfm.de`

Ein Nachteil der sich durch die informelle Natur der Verschlagwortung ergibt ist, dass die entstehenden Begriffssysteme inkonsistent werden. Hier sind Synonyme als typisches Beispiel zu nennen. Ein Artikel über die Krankheit Alzheimer könnte durch einen Nutzer als "Alzheimer" verschlagwortet werden wohingegen ein anderer Nutzer eher "Demenz" als Schlagwort verwendet. Neben Synonymen ist auch die Polysemie ein Problem. So können Inhalte die thematisch nicht verwandt sind trotzdem gleich verschlagwortet sein (Firma Apple - Obst Apple).
Eine Möglichkeit die Inkonsistenz der Begriffssysteme zu verringern ist die Nutzung eines kontrollierten Vokabulars, wie das Unified Medical Language System (UMLS). Ein kontrolliertes Vokabular diktiert eine Menge von Begriffen die bevorzugt werden sollen, um dann von den Nutzern ausgewählt zu werden. Da bei der Formulierung einer Suchanfrage ebenfalls auf das kontrollierte Vokabular zurückgegriffen wird eliminiert man die Diskrepanz zwischen Begrifflichkeiten (IJzereef et al., 2005). Es gilt nun in Erfahrung zu bringen wie ein System eine semi-automatische Verschlagwortung von Inhalten auf Basis eines kontrollierten Vokabulars bieten kann, um auf diese Weise den Nutzer bei der Auswahl geeigneter Schlagwörter zu unterstützen. Das Verwenden eines kontrollierten Vokabulars bietet zahlreiche vielversprechende Vorteile. In einem ersten Schritt unterstützt das kontrollierte Vokabular die Nutzer bei der Verschlagwortung von Inhalten. Es agiert dabei jedoch lediglich als Mediator, der Vorschläge unterbreitet statt diese gezielt zuzuweisen. Auf diese Weise behält die Verschlagwortung ihre informelle Natur, wird jedoch durch das kontrollierte Vokabular angereichert. Ein weiterer Vorteil der Nutzung eines kontrollierten Vokabulars ist die Möglichkeit externe Inhalte, die auf der selben Terminolgie aufbauen in das System zu integrieren, wodurch ein System mit relevanten externen Informationen angereichert werden kann.

5.1 Vorgehen

Der Prozess der automatischen Verschlagwortung ist in Abbildung 5.2 dargestellt. Im Wesentlichen lässt sich die automatische Verschlagwortung in zwei Phasen unterteilen. Zunächst erfolgt eine Vorverarbeitung des zu analysierenden Dokuments die in drei Teilprozesse unterteilt ist. In einem ersten Schritt erfolgt die Zerlegung des Dokuments in wortähnliche Einheiten (Token). Die Teilprozesse *Stemming* und *POS-Tagging* dienen der Filterung des Dokuments, um die Menge an Begriffen zu verringern, welche mit dem UMLS Metathesaurus analysiert werden. Dies geschieht indem Token auf ihren Wortursprung transformiert (Stemming) und bestimmte, zuvor definierte grammatikalische Elemente ignoriert werden (POS-Tagging). Die Relevanzbewertung der verbliebenen Token erfolgt während der Merkmalsanalyse unter Verwendung des UMLS Metathesaurus. Anschließend werden dem Nutzer die Ergebnisse nach Termrelevanz geordnet präsentiert. Im Folgenden wird die Vorverarbeitung sowie die Merkmalsanalyse im Detail erläutert. Das vorgestellte Prozessmodell lässt sich verallgemeinern und auf eine domänenübergreifende Verschlagwortung textbasierter Inhalte anwenden. Voraussetzung dafür ist das Vorhandensein eines domänenspezifischen Begriffssystems.

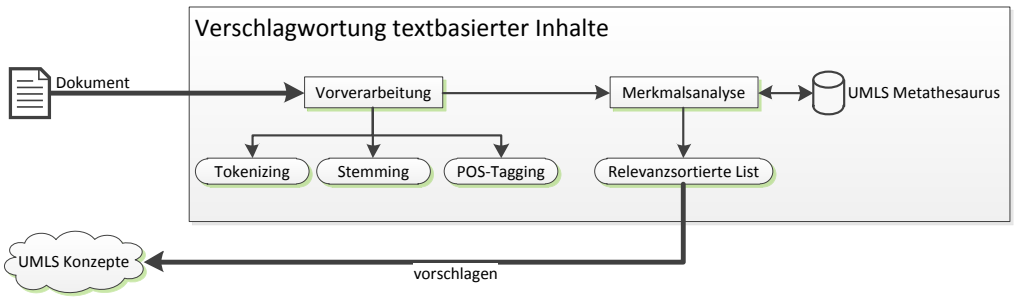


Abbildung 5.2: Prozesse der automatischen Verschlagwortung unter Verwedung eines Begriffssystems

5.2 Tokenizing textbasierter Inhalte

Um eine Analyse textbasierter Inhalte vorzunehmen müssen zunächst eine Reihe von Vorverarbeitungsprozessen durchgeführt werden. Ein solcher Vorverarbeitungsprozess

ist das *Tokenizing*, bei dem es sich um die Isolation von wortähnlichen Einheiten (*Token*) in einem Text handelt. In der Regel wird dies durch zwei Transformationen erreicht. In der ersten Transformation werden Sätze aus dem Originaltext isoliert. In der zweiten Transformation werden die isolierten Sätze wiederum in einzelne Token zerlegt (Grefenstette und Pasi, 1994). Dabei werden je nach Anwendungsgebiet Regeln festgelegt die definieren, welche Muster der Erkennung von Token-Grenzen dienen und wie mit Sonderzeichen umzugehen ist. Das Tokenizing mag an dieser Stelle trivial erscheinen, jedoch ist es ein entscheidender Prozess bei der Analyse von textbasierten Inhalten, da es als Fundament für sämtliche Folgeprozesse dient. In (Klatt und Bohnet, 2005) beschreiben die Autoren das Tokenizing als den einzigen Prozess, der sich in sämtlichen Anwendungen wiederfindet, die sich mit der Verarbeitung von natürlicher Sprache (*Natural Language Processing*) beschäftigen und der grundsätzlich auch der erste Schritt einer jeden Analyse ist.

Ein naheliegender Ansatz beim Tokenizing ist die Verwendung von Leer- und Satzzeichen als Token-Grenze. Dies soll nun an folgendem Beispieltext verdeutlicht werden.

> Herr Stadtler fuhr heute mit dem Fahrrad zur Arbeit. Auf dem Weg traf er seine alte Schulfreundin Jana.

Tokenizer, die Leer- und Satzzeichen als Token-Grenze verwenden können den oben dargestellten Text problemlos in zwei Sätze und 18 Token unterteilen. Dieser Ansatz mag für eine Vielzahl von Texten zu befriedigenden Ergebnissen führen stösst jedoch bei komplexeren Inhalten schnell an seine Grenzen. Typische Beispiele für eine unzureichende Verwendung von Leer- und Satzzeichen als Token-Grenze sind:

- Satzzeichen: *Dr. etc. U.S.A. 1. Februar 2011*
  Aufzählungen, Abkürzungen und Datumsangaben beinhalten in der Regel Punkte welche in dem Fall keine Token-Grenzen darstellen.

- Apostrophe: *Das Motto von Klaus war immer "Ende gut alles gut"*
  Das Apostroph als Kennzeichnung eines Zitats würde zu den fehlerhaften Token *"Ende* und *gut"* führen.

- Satzzeichen : *Wuppertal,Köln*
  Das Fehlen von Leerzeichen zwischen Satzanfang und Satzende würde zu dem Token *Wuppertal.Köln* führen

Neben naiven Ansätzen wie dem Tokenizing unter Verwendung von Leer- und Satzzeichen existieren auch komplexere Methoden wie die Klassifikation, nichtüberwachte Klassifikation und Heuristiken.

Ein Vorteil der vorliegenden Arbeit ist die Tatsache, dass die Domäne bereits bekannt ist. Dadurch wird es möglich bei Prozessen wie dem Tokenizing domänenspezifische Methoden zu verwenden. Dies wird im folgenden Abschnitt näher erläutert.

5.2.1 Domänenspezifischer Tokenizer

Neben den bereits aufgezeigten Eigenschaften stellt das Tokenizing von medizinischen Texten eine zusätzliche Herausforderung dar (Barrett und Weber-Jahnke, 2009) , (Tomanek et al., 2007). Tokenizer, dessen Regeln zur Zerlegung auf einfachen Vorgaben wie Leerzeichen basieren liefern bei Inhalten aus der Medizin oft unbefriedigende Ergebnisse. Dies kann an einem Beispiel aus der Biomedizin verdeutlicht werden. In der Biomedizin stellt *CD28-dependent* eine Verbindung zwischen Entitäten dar, welche in zwei Token resultieren sollten, namlich *CD28* und *dependent* (Tomanek et al., 2007). Die Problematik des Tokenizing medizinischer Inhalte wird ebenfalls in (Jiang und Zhai, 2007) behandelt. Dabei werden verschiedene Heuristiken vorgestellt mit dessen Hilfe ein Tokenizing von medizinischen Inhalten vorgenommen werden kann. Die dort beschriebenen Heuristiken lauten:

- Ersetzen von !"#$%&∗ <=>?@\| ∼ durch Leerzeichen
- Entfernen von . :; , wenn darauf ein Leerzeichen folgt
- Entfernen von ()[] wenn vor der öffnenden Klammer ein Leerzeichen steht und auf die schliessende Klammer ein Leerzeichen folgt
- Entfernen von $'$ wenn davor oder danach ein Leerzeichen steht
- Entfernen von $'s$ und $'t$ wenn darauf ein Leerzeichen folgt
- Entfernen von / wenn darauf ein Leerzeichen folgt

Es zeigt sich, dass durch die speziellen Eigenschaften der medizinischen Begriffsbildung die Ansprüche an entsprechende Algorithmen steigen. Mittlerweile existieren jedoch Tokenizer, die speziell für den medizinischen Bereich entwickelt sind. Einige dieser Tokenizer werden im technischen Report von (He und Kayaalp, 2006) in Hinblick auf ihre Einsatzmöglichkeiten im Kontext der Medizin verglichen. Insgesamt vergleichen die Au-

toren 13 Tokenizer, welche sich im Wesentlichen in der Definition von Token-Grenzen (Trennung bei Leerzeichen) und dem Umgang mit Aufzählungen bzw. Zahlenfolgen unterscheiden. Um die Arbeitsweise der Tokenizer zu simulieren wurden verschiedene Beispieltexte entwickelt. Einer der analysierten Beispieltexte ist in Abbildung 5.3 dargestellt. Der Beispieltext mit dessen Hilfe die Funktionsweise der Tokenizer simuliert

> Independent of current body composition, IGF-I levels at 5 yr were significantly associated with rate of weight gain between 0-2 yr (beta = 0.19; P < 0.0005), and children who showed postnatal catch-up growth (i.e. those who showed gains in weight or length between 0-2 yr by >0.67 SD score) had higher IGF-I levels than other children (P = 0.02).

Abbildung 5.3: Beispieltext zur Analyse der 13 Tokenizer aus (He und Kayaalp, 2006)

wird enthält Interpunktionszeichen, mathematische Symbole und auch Bindestriche.

Aufbauend auf den Ergebnissen von (He und Kayaalp, 2006) wird nun ermittelt, welche Tokenizer für die weitere Verarbeitung in Frage kommen. Dabei wurde auf Tokenizer verzichtet, die sich entweder nicht auf der verwendeten Entwicklungsumgebung installieren ließen (Mac OS X 10.6.8) oder nicht mehr verfügbar waren. Weiterhin produzierten einige Tokenizer identische Ergebnisse, weshalb in einem solchen Fall nur einer der Tokenizer verwendet wurde. Ein Tokenizer wurde ignoriert, da er unbrauchbare Ergebnisse produzierte indem beispielsweise Zahlen und Klammern aus dem Text entfernt werden. Dies wird deutlich, wenn man den Beispieltext aus Abbildung 5.3 mit dem in Abbildung 5.4 dargestellten Ergebnis von Tokenizer 3 vergleicht (Token sind durch ein " | " getrennt).

> Independent | of | current | body | composition | IGF | levels | at | yr | were | significantly | associated | with | rate | of | weight gain | between | yr | beta | P | lt | and | children | who | showed | postnatal | catch | up | growth | i | e | those | who | showed | gains | in | weight | or | length | between | yr | by | SD | score | had | higher | IGF | I | levels | than | other | children |

Abbildung 5.4: Erzeugte Token eines Beispieltextes für Tokenizer 3

Neben den verbliebenen Tokenizern aus dem technischen Report kommt zusätzlich der Tokenizer *JULIE TBD (Jena University Language Information Engineering Lab - Token Boundary Detector)* zum Einsatz[6]. Die vollständige Liste der insgesamt neun

[6] `https://julielab.de/Resources/Software/NLP_Tools.html`

Nr.	Tokenizer
1	NLTK Tokenizer
2	OpenNLP Tokenizer
3	SPECIALIST NLP Tokenizer
4	Dan Melamed's Tokenizer
5	UIUC word splitter
6	LT TTT Tokenizer
7	MedPost Tokenizer
8	Stanford POS tagger
9	JTBD

Tabelle 5.1: Liste der analysierten Tokenizer

Tokenizer findet sich in Tabelle 5.1. Diese werden nun hinsichtlich der Einsatzmöglichkeiten als Tokenizer für das vorliegende Projekt getestet. Das Hauptaugenmerk liegt dabei auf dem Tokenizing von Konzeptbezeichnungen aus dem UMLS Metathesaurus.

5.2.2 Tokenizer im UMLS Metathesaurus

In Hinblick auf die Verwendung des UMLS Metathesaurus als medizinisches Begriffssystem sollen nun zunächst einige typische Wortkonstrukte dargestellt werden, die sich innerhalb des UMLS Metathesaurus finden. Einige komplexere Wortkonstrukte sind in Tabelle 5.2 dargestellt.

CUI	Bezeichnung
C0001476	Ca(2+)-Transport-ATPase
C0886672	Ondansetron, (+-)-Isomer
C0001964	1-Propanol
C0000039	1,2-Dipalmitoylphosphatidylcholin
C0002395	Alzheimer-Krankheit
C0969358	Kv1.2'-Kanal
C0161398	Verletzungen des N. opticus
C0023896	Fettleber, alkoholische

Tabelle 5.2: Spezielle Wortkonstrukte im UMLS Metathesaurus

Die UMLS Beispielkonzepte werden nun als Eingangstext für die neun zu analysierenden Tokenizer verwendet. In Tabelle 5.3 sind die erzeugten Token für die UMLS Beispielkonzepte dargestellt. Es ist zu erkennen, dass sich teils deutliche Unterschiede in der Menge der erzeugten Token finden. Dies wird deutlich, wenn man die Gesamt-

menge an erzeugten Token für die 8 Beispielkonzepte in Abbildung 5.5 betrachtet. Auffällig sind hier der Tokenizer 2 mit lediglich 19 Token und Tokenizer 3 mit insgesamt 42 Token. Ein wichtiger Aspekt ist das Verhalten von Tokenizern bei Begriffen, die durch einen Bindestrichen verknüpft sind. Dabei können Tokenizer solche Konzepte entweder als ein Token verarbeiten oder eine Trennung vornehmen und mehrere Token erzeugen. Das Konzept Alzheimer-Krankheit wird von den Tokenizern 1, 5, 6, 7 und 8 als ein Token behandelt wohingegen die Tokenizer 3, 4 und 9 drei Token (*Alzheimer* | - | *Krankheit*) erzeugen. Bei Tokenizer 9 ist zu beobachten, dass ein Bindestrich nicht zwangsläufig zu einer Trennung der Begriffe führt, wie man an dem Konzept *1,2-Dipalmitoylphosphatidylcholin* erkennen kann. Auch wenn hier ein Bindestrich im Konzept zu finden ist erzeugt Tokenizer 9 hier lediglich einen Token.

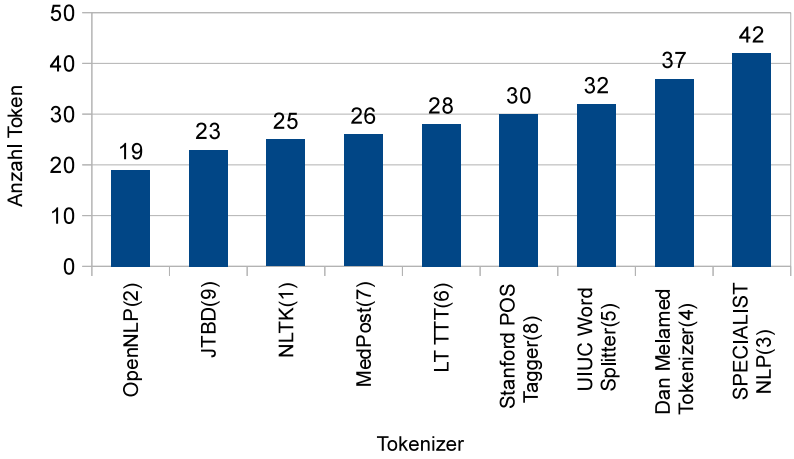


Abbildung 5.5: Anzahl erstellter Token je Tokenizer

5.2.3 Bewertung der Tokenizer

Grundsätzlich muss vor der Entscheidung für einen bestimmten Tokenizer festgelegt werden welche Ergebnisse für das System wünschenswert sind. Hier wird zwischen der Granularität des Systems unterschieden. Ein Tokenizer, der Inhalte besonders feingranular verarbeitet erzeugt kürzere Token, was gleichzeitig eine grössere Gesamtmenge an Token bedeutet. Analog dazu bedeutet dies, dass bei einem grobgranularen Tokenizer längere Token enstehen und sich gleichzeitig die Gesamtmenge an Token verringert. Dies soll an dem Beispiel *Kv1.2'-Kanal* verdeutlicht werden. Die Tokenizer 1, 2 und 8 verarbeiten den Begriff unverändert und erzeugen eine Tokenmenge mit lediglich einem

Variation	Tokenizer	#Token
Ca(2+)-Transport-ATPase		
Ca(2+)-Transport-ATPase	6,9	1
Ca(| 2+ |)-Transport-ATPase	2	3
Ca | (| 2+ |) | -Transport-ATPase	7	5
Ca | (| 2 | + |) | -Transport-ATPase	1	6
Ca | (| 2 | + |) | - | Transport-ATPase	5,8	7
Ca | (| 2 | + |) | - | Transport | - | ATPase	3,4	9
Ondansetron,(+-)-Isomer		
Ondansetron | , | (| +-)-Isomer	2	4
Ondansetron | , | (+-) | - | Isomer	9	5
Ondansetron | , | (| +- |) | -Isomer	7	6
Ondansetron | , | (| + | - |) | -Isomer	1	7
Ondansetron | , | (| + | - |) | - | Isomer	3,4,5,6,8	8
1-Propanol		
1-Propanol	1,2,5,7,8	1
1 | - | Propanol	3,4,6,9	3
1,2-Dipalmitoylphosphatidylcholin		
1,2-Dipalmitoylphosphatidylcholin	1,2,5,8,9	1
1,2 | - | Dipalmitoylphosphatidylcholin	4,6	3
1 | , | 2 | - | Dipalmitoylphosphatidylcholin	3,7	5
Alzheimer-Krankheit		
Alzheimer-Krankheit	1,2,5,6,7,8	1
Alzheimer | - | Krankheit	3,4,9	3
Kv1.2'-Kanal		
Kv1.2'-Kanal	1,2,7	1
Kv1.2' | - | Kanal	4,9	3
Kv | 1.2 | ' | - | Kanal	5	5
Kv1 | .2 | ' | - | Kanal	6,8	5
Kv1 | . | 2 | ' | - | Kanal	3	6
Verletzungen des N. opticus		
Verletzungen | des | N. | opticus	5,6,7,8,9	4
Verletzungen | des | N | . | opticus	1,2,3,4	5
Fettleber, alkoholische		
Fettleber | , | alkoholische	1,2,3,4,5,6,7,8,9	3

Tabelle 5.3: Ergebnisse der Tokenizer für die UMLS Beispielkonzepte

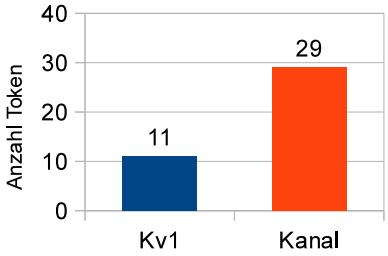


Abbildung 5.6: Suche nach den Token Kv1 und Kanal im UMLS Metathesaurus

Token: $T = \{Kv1.2' - Kanal\}$. Im Vergleich dazu erzeugt der Tokenizer 3 eine Tokenmenge mit sechs Token: $T = \{Kv1 \mid . \mid 2 \mid ' \mid - \mid Kanal\}$ was als feine Verarbeitung bezeichnet werden kann.

Die Auswirkungen auf die Qualität des Systems sind dabei gravierend. Man kann davon ausgehen, dass bei feinerem Tokenizing erheblich mehr Inhalte identifiziert werden können, die Präzision der Ergebnisse jedoch darunter leidet. Dies lässt sich durch eine einfache SQL Abfrage an die UMLS Metathesaurus Datenbank verdeutlichen. Findet sich ein Token in der Datenbank, so wird davon ausgegangen, dass es sich um einen relevanten medizinischen Begriff handelt. Ein Abgleich zwischen Token und der MySQL Datenbank des UMLS Metathesaurus erfolgt über die Suche nach übereinstimmenden Zeichenfolgen am Anfang und Ende einer jeden Konzept-Bezeichnung. Weiterhin werden Token ignoriert, die eine Länge von 3 Zeichen unterschreiten. Unter diesen Voraussetzungen lautet die Menge der zu analysierenden Token:

- Tokenizer 1,2 und 8
  $T = \{Kv1.2' - Kanal\}$
- Tokenizer 3
  $T = \{Kv1|Kanal\}$

Für die Tokenizer 1, 2 und 8 die lediglich einen zu analysierenden Token ermittelt haben ist die Ergebnismenge der SQL Abfrage erwartungsgemäß klein. Es findet sich lediglich das Konzept *C0969358 Kv1.2'-Kanal* im UMLS Methathesaurus. Wie in Abbildung 5.6 zu erkennen ist, ist die Ergebnissmenge für den Tokenizer 3 ungleich größer. Für den Token *Kv1* finden sich 11 Konzepte und für den zweiten Token *Kanal* sind es 29 Konzepte.

Einer der Vorteile der Verwendung eines feingranularen Tokenizer ist die Möglichkeit auch bei geringen Abweichungen in der Schreibweise das richtige Konzept zu finden. Angenommen ein Nutzer schreibt Kv1.2-Kanal statt Kv1.2'-Kanal, so würde ein grobgranularer Tokenizer das passende Konzept im UMLS Metathesaurus nicht finden wohingegen bei Verwendung eines feingranularen Tokenizer auch die geringe Abweichung in der Schreibweise das passende Konzept in der Ergebnismenge liefern würde. Dennoch ist der Berechnugsaufwand bei Verwendung eines feingranularen Tokenizer zu groß, so dass für die weitere Verarbeitung ein grobgranularer Tokenizer wünschenswert ist. Dies ist so lange der Fall wie keine Eingrenzung der Inhalte im UMLS Metathesaurus durch Domänenexperten vorgenommen wird. Zusammenfassend lässt sich festhalten, dass ein Tokenizer wünschenswert ist, der eine möglichst geringe Menge an erzeugten Token aufweist. Tokenizer die feingranular arbeiten erzeugen beim Abgleich mit dem UMLS Metathesaurus zuviel Rauschen und liefern Ergebnisse mit unzureichender Präzision. Dies lässt sich an dem Tokenizer 3 und dem Beispielkonzept Kv1.2-Kanal verdeutlichen. Tokenizer 3 trennt Begriffe bei dem Auftreten eines Bindestrichs in einzelne Token. Auf diese Weise ensteht der Token Kanal, welcher bei Abgleich mit dem UMLS Metathesaurus zwar zahlreiche medizinisch relevante Ergebnisse liefert aber eben auch Inhalte die medizinisch nicht relevant sind. Exemplarisch steht dafür das Konzept *C0007950 Kanalinseln* was keinen medizinischen Begriff beschreibt sondern eine Inselgruppe die im UMLS unter der Kategorie *Geographics* zu finden ist. Eine verkürzte Liste an gefundenen Konzepten für Tokenizer 3 ist in Tabelle 5.4 dargestellt. Eine vollständige Auflistung der gefundenen Konzepte findet sich im Anhang. 8.4. Unter Berücksichti-

Token	CUI	Bezeichnung
Kv1	C0290730	Kv1.3-Kaliumkanal
Kv1	C0294031	Kv1.1-Kaliumkanal
Kv1	C0299095	Kv1.4-Kaliumkanal
..	..	..
Kanal	C0007950	Kanalinseln
Kanal	C0917707	Kanalisierung
Kanal	C0910872	Zyklisch-Nukleotid-gesteuerter Kationenkanal
Kanal	C0969358	Kv1.2'-Kanal
..	..	..

Tabelle 5.4: Suche nach den Token Kv1 und Kanal im UMLS Metathesaurus

gung der Ergebnisse aus Abbildung 5.5 kommen für den weiteren Verarbeitungsprozess die Tokenizer 2 und 9 in Frage. Beide Tokenizer erzeugen eine geringe Menge an Token und verhindern somit, dass Begriffe fälschlicherweise als medizinisch relevant betrachtet werden.

5.2.4 Tokenizer Implementation

Im vorangegangenen Abschnitt wurde aufgezeigt, dass für Projekte die den UMLS Metathesaurus als Begriffssystem einsetzen ein grobgranularer Tokenizer wünschenswert wäre. Die beiden Tokenizer mit der geringsten Menge an Token sind die der OpenNLP (Tokenizer 2) und JULIE Token Boundary Detector (Tokenizer 9). OpenNLP und JTBD werden nun kurz erläutert und im Anschluss werden beide Systeme anhand von fünf Beispieltexten miteinader verglichen.

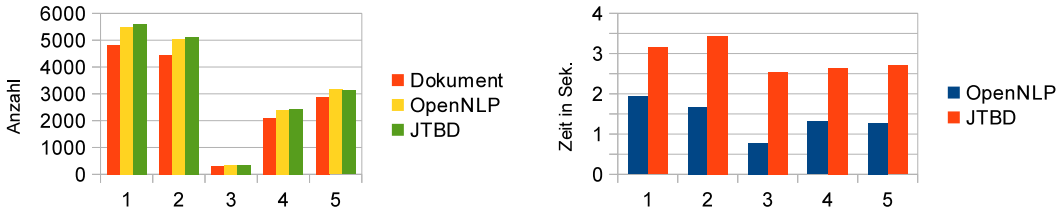
OpenNLP

Bei OpenNLP[7] handelt es sich um ein Werkzeug zur Verarbeitung von textuellen Inhalten. Ursprünglich wurde OpenNLP von Jason Baldridge und Gann Bierner an der Universität Edinburgh entwickelt und wechselte im Jahr 2010 zu einem Apache Incubator Projekt[8]. OpenNLP wurde in Java entwickelt und befindet sich aktuell[9] in der Version 1.5.2 .
Beim Tokenizing verwendet OpenNLP machine learning Algorithmen und stellt zahlreiche bereits trainierte Modelle in verschiedenen Sprachen, darunter auch Deutsch, unter `http://opennlp.sourceforge.net/models-1.5/` zur Verfügung. Neben vordefinierten Modellen erlaubt OpenNLP auch das Antrainieren eines eigenen Modells um auf diese Weise das Tokenkizing anzupassen.

JTBD

Bei JTBD (Jena University Language Information Engineering Lab - Token Boundary Detector) handelt es sich um einen Tokenizer, der speziell für den biomedizinischen Bereich entwickelt und optimiert wurde. Betrieben wird JTBD von der Universität Jena - Labor für Sprachen und Informationen[10]. Zu den Funktionalitäten von JTBD zählt das Tokenizing, das Trainieren eines eigenen Modells und die Evaluation (Tomanek, 2007). JTBD wurde in Java entwickelt und befindet sich zur Zeit[11] in der Version 1.6. Zum Lieferumfang von JTBD zählt ein bereits trainiertes Modell, welches durch

[7] `http://incubator.apache.org/opennlp/`
[8] Bei Apache Incubator `http://incubator.apache.org/` handelt es sich um einen Einstiegspunkt in die Apache Software Foundation. Alle Projekt die der Apache Software Foundation beitreten wollen durchlaufen zunächst eine Phase als Incubator Projekt (Apache Incubator, 2012)
[9] Stand 01/2012
[10] `https://julielab.de/JULIE+Lab.html`
[11] Stand 01/2012



(a) Anzahl erzeugter Token

(b) Verarbeitungsgeschwindigkeit der Tokenizer

Abbildung 5.7: Ergebnisse der Tokenizer für 5 Testdokumente

einen sequentiellen Algorithmus aus dem Bereich des maschinellen Lernens erzeugt wurde (Tomanek et al., 2007). Als Trainingsdaten kommen Journalzusammenfassungen aus dem biomedizinischen Bereich zum Einsatz, welche durch zwei Plattformen zur Verfügung gestellt werden (GENIA[12] und PennBioIE[13]).

Vergleich zwischen openNLP und JTBD

Um zu ermitteln, welcher der beiden Tokenizer geeignet ist, wurden beide Systeme mit fünf Beispieltexten getestet. Dabei handelt es sich um einen Online-Blogeintrag zum Thema Leberzirrhose, sowie vier Zeitschriftenbeiträge aus dem Journal *Der Chirurg*. Beide Tokenizer wurden innerhalb der Kommandozeile mit den mitgelieferten, auf die Deutsche Spache ausgelegten Modellen ausgeführt. Die Ergebnisse der Tokenizer für die fünf Beispieltext sind in Abbildung 5.7 dargestellt.

Wie zu erwarten war erzeugen beide Tokenizer mehr Token als die anfänglich über Leerzeichen ermittelte Zahl an Wörtern in dem Text. Bis auf zwei Ausnahmen (Dokument drei und fünf) erzeugt JTBD mehr Token als der OpenNLP, was über die Gesamtmenge an Token eine Differenz von 151 Token ergibt. Dieser Unterschied ist in dem Fall nicht gravierend. Einen deutlichen Unterschied sieht man bei Betrachtung der Laufzeit. Über die Gesamte Menge an Dokumenten benötigt OpenNLP 6,92 Sekunden und ist somit deutlich schneller als JTBD mit 14,52 Sekunden.

Hinsichtlich einer Implementation weisen beide Systeme kaum Unterschiede auf. Beide wurden in Java entwickelt und erlauben das Trainieren eines eigenen Modells. Unterschiede finden sich jedoch bei der Verarbeitung der fünf Beispieltexte. Aufgrund der

[12] `http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/`
[13] `http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2008T21`

deutlichen Unterschiede in der Laufzeit und der geringeren Zahl an erzeugten Token die aus den Beispieltexten erstellt wurden wird für den weiteren Prozess der automatischen Verschlagwortung der OpenNLP Tokenizer verwendet.

5.3 Filterung

Der vorangegangene Abschnitt hat gezeigt wie Text in einzelne Token zerlegt werden kann. In Abhängigkeit des Anwendungsgebiets erfolgt während der Verarbeitung natürlicher Sprache üblicherweise eine Merkmalsanalyse bei dem die ermittelten Token mit weiteren Eigenschaften angereichert werden. Das wohl bekannteste Merkmal eines Token ist die Häufigkeit mit der dieser in einem Text auftritt. Um jedoch die Häufigkeit eines Token zuverlässig zu ermitteln muss ein System in der Lage sein Flexionsformen auf eine gemeinsame Stammform zu transformieren (Lampen ⇒ Lampe). Zu diesem Zweck folgt in diesem Abschnitt eine Analyse von Token-Merkmalen, die sich aus zwei Teilprozessen zusammensetzt: dem *Stemming* bei dem Token auf ihre Stammform transformiert werden und dem *Part-of-speech Tagging* (POS-Tagging) bei dem jeder Token einer grammatikalischen Kategorie zugewiesen wird. Ein weiterer Hintergrund aus dem die beiden hier vorgestellten Prozesse eingesetzt werden ist die Reduktion von Token, die mit dem UMLS-Metathesaurus verglichen werden sollen. Eines der Ziele dieser Arbeit ist die Identifikation medizinisch relevanter Begriffe in einem Text unter Verwendung eines Begriffssystems. Betrachtet man jedoch die Menge an Token, die sich aus einem Text ergeben wird schnell deutlich, dass ein Abgleich mit dem Begriffssystem zu einer unüberschaubar grossen Ergebnissmenge führen würde. Daher dient der folgende Abschnitt zudem der Filterung von Token. Es ist hier wichtig anzumerken, dass eine Filterung nicht auf ein Entfernen der Token abzielt sondern lediglich Eingrenzen soll, welche Token mit dem UMLS-Metathesaurus verglichen werden sollen.

5.3.1 Stemming

Bei der Analyse von natürlicher Sprache ist es ein wichtiger Aspekt zu ermitteln, wie häufig ein bestimmtes Wort in einem Text vorkommt. Dabei gilt die Annahme, dass Wörter die häufiger in einem Text erwähnt werden auch von grösserer Bedeutung sind. Es wird schnell deutlich das ein einfaches Vergleichen der verschiedenen im Text enthaltenen Wörter nicht ausreicht, um eine qualitative Aussage über die Häufigkeit von Begriffen zu treffen. Dies ist auf die sogenannte Flexion zurückzuführen bei der die Schreibweise eines Wortes von seiner Funktion im Satz abhängt. Typische Beispiele für eine Flexion sind "Monitor" ⇒ "Monitore" ⇒ "Monitors" oder "rennst" ⇒ "ranntest" ⇒ "gerannt", welche sich auf eine gemeinsame Stammform zurückführen lassen (Gottron, 2010). Die Rückführung einer Flexion auf seine Stammform stellt Systeme, die eine maschinelle Verarbeitung natürlicher Sprache durchführen vor eine Herausforderung. Eine Möglichkeit, Worte auf ihre Stammform zurückzuführen ist das soge-

nannte *Stemming*, bei dem der Suffix eines Wortes anhand von vordefinierten Reglen eliminiert wird. Der verleibende Teil des Wortes wird dann als der sogenannte *Stem* bezeichnet. Neben dem Stemming existieren weitere Methoden, die der Erkennung von Wortgemeinsamkeiten dienen. Typische Einsatzbereich sind das Erkennen von orthografischen Fehlern (Kaffemaschine), Tippfehlern (Kaffeeamschine) und Formulierungsvarianten (Kaffee-Vollautomat). Hier kommen Ähnlichkeitsmaße zum Einsatz, wie die *Trigramm-Ähnlichkeit* und die *Edit-Distance* die ermitteln wie ähnlich sich zwei Zeichenketten sind (Reichenberger, 2010). Für die vorliegende Arbeit wird jedoch eine Beschränkung auf Stemmer vorgenommen. Einer der populärsten Stemming-Algorithmen ist der *Porter-Stemmer*, welcher 1980 von Porter entwickelt wurde (Porter, 1980) und ebenfalls auf der Annahme basiert, dass ein Suffix sich in der Regel auf der rechten Seite eines Begriffs befindet. Ursprünglich wurde der Porter-Stemmer für die englische Sprache entwickelt. Mittlerweile finden sich jedoch unter dem *Snowball Projekt*[14] auch Stemmer für zahlreiche weitere Sprachen, darunter auch ein Stemmer für die deutsche Sprache wieder.

Für die Suche nach medizinisch relevanten Token im UMLS Metathesaurus spielt das Stemming eine wichtige Rolle, da es verhindern soll, dass Token mehrfach analysiert werden. Dies führt zum einen zu einer Steigerung der Performanz und zum anderen dazu, dass zu sämtlichen Flexionen eines Token nur genau ein Relevanzwert berechnet wird. In Abbildung 5.8 sind die Ergebnisse des Stemming für zwei Beispielbegriffe dargestellt. Es ist zu erkennen, dass sich fünf Token auf die Stammform "arteriell" stemmen lassen. Es wird jedoch auch ersichtlich, dass Stemming zu einer Verfälschung der Token führen kann, die zur Folge hat das Token die von unterschiedlicher Bedeutung sind auf eine gemeinsame Stammform transformiert werden. Bei dem Token "Eisen" beispielsweise ist das Element Eisen gemeint was nach dem Stemming zu "Eis" transformiert wurde, was gleichzeitig die Stammform von "Eis" ist.

Es ist wichtig anzumerken, dass die Suche nach passenden Konzepten im UMLS Metathesaurus weiterhin auf Basis der Token vor dem Stemming stattfindet. Aus diesem Grund können kleinere Fehler die beim Stemming entstehen (Eisen ⇒ Eis, Eis ⇒ Eis) ignoriert werden, da diese lediglich zu einer geringen Verfälschung der Worthäufigkeit führen. Bei Token die mehrmals im Text vorkommen und bei Token die sich zwar in ihrer Schreibweise unterscheiden aber durch Stemming auf eine gemeinsame Stammform gebracht wurden wird nur einmal der UMLS Referenzwert errechnet.

[14] `http://snowball.tartarus.org/`

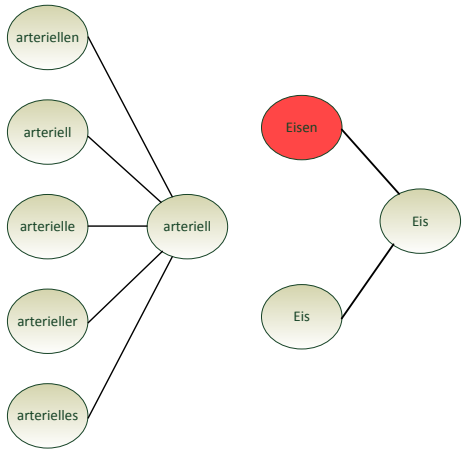


Abbildung 5.8: Ergebnisse des Stemming für die Begriffe arteriell und Eisen

5.3.2 POS-Tagging

Nachdem nun bereits eine Reduktion der zu analysierende Token durch Stemming erreicht wurde, folgt nun eine Beschränkung auf bestimmte grammatikalische Tokeneinheiten. Bei der automatischen Verschlagwortung des Eingangstextes wird deutlich, dass auch nach dem Stemming zahlreiche weitere Token existieren die sich für eine Suche im UMLS Metathesaurus nur bedingt eignen. Betrachtet man den UMLS Metathesaurus ist zu erkennen, dass Konzeptbezeichnungen primär durch ein Substantiv beschrieben werden. Weitere Fälle sind Substantive in Kombination mit Adverben (*C0015696 Fettleber, alkoholisch*) oder auch Teilsätze wie bspw. *C0008078 Verhalten des Kindes* . Eine Konzeptbezeichnung ohne Substantiv besitzt nur geringe Aussagekraft für ein Konzept. Aus diesem Grund ist es sinnvoll das Auffinden von passenden Konzepten auf Token die ein Substantiv darstellen zu beschränken. Dies führt zu einer erheblichen Verbesserung der Performanz und dazu, dass weniger Rauschen in den Ergebnissen zu finden ist. Durch die Übereinstimmungskriterien die zum Abgleich mit dem UMLS Metathesaurus gewählt wurden bleibt dennoch sichergestellt, dass auch Konzepte identifiziert werden die aus mehr als einem Substantiv bestehen.
Um eine Beschränkung auf Substantive bei der Suche im UMLS Metathesaurus zu erreichen wird eine Möglichkeit benötigt diese in einem Text zu identifizieren. Dies kann durch Verwendung von sogenannten *Part-of-Speech Taggern* erreicht werden die jedem Token in einem Text eine grammatikalische Kategorie zuweisen. In der Regel erfolgt die Zuweisung grammatikalischer Kategorien auf Basis einer vollständigen syntaktischen Zerlegung, bei der eine sprachliche Beziehung von einem Token zu allen anderen

Token in einem Text hergestellt wird. Im Gegensatz zu der Verwendung einfachere Wörterbuch-Ansätze ermöglicht dies die Auflösung von Mehrdeutigkeiten und erhöht die Präzision der Ergebnisse (Granitzer, 2006). Die Entwicklung von POS-Taggern fand ursprünglich im englischen Sprachraum statt, doch existieren mittlerweile POS-Tagger die auch die deutsche Sprache unterstützen. Ein POS-Tagger der zahlreiche Sprachen unterstützt ist der 1995 von Helmut Schmid an der Universität Stuttgart entwickelte *TreeTagger*[15]. Entwickelt wurde der TreeTagger auf Basis von *Markov-Modellen* und erzielt beim POS-Tagging eine Genauigkeit von 97% (Schmid, 1995). Da der TreeTagger in C entwickelt wurde kommt ein sogenannter *TreeTagWrapper*[16] zum Einsatz der eine Nutzung auch aus der Programmiersprache Java heraus erlaubt. In Abbildung 5.9 sind die Ergebnisse des TreeTagger für einen Beispielsatz aus den Trainingsdaten dargestellt. Es ist zu erkennen das jedem Token in dem Satz ein ent-

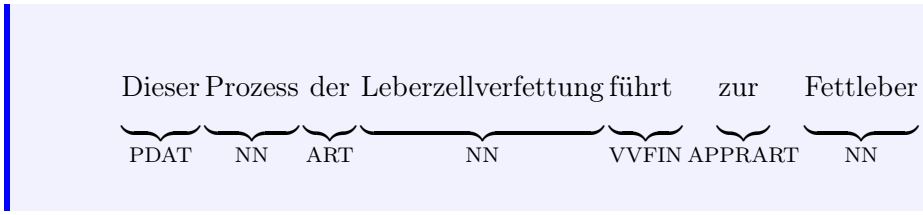
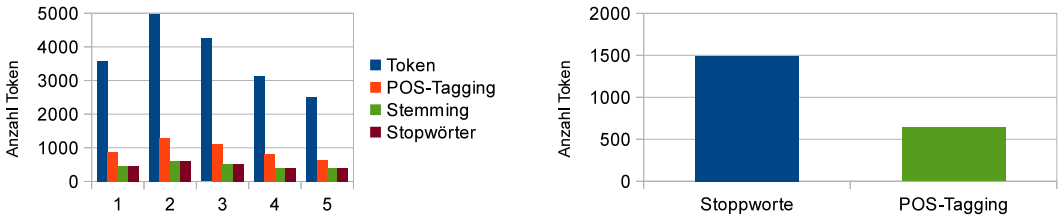


Abbildung 5.9: Ergebnisse des TreeTagger zu einem Beispieltext

sprechender Tag zugewiesen wird, der aussagt um welche grammatikalische Kategorie es sich bei dem Token handelt. Eine vollständige Liste des verwendeten Tagsets ist unter `ftp://ftp.ims.uni-stuttgart.de/pub/corpora/stts_guide.pdf` erhältlich. In dem Beispieltext finden sich drei Substantive alle mit dem Tag "NN" (normale Nomina) versehen. Weitere Tags in dem Beispieltext sind "PDAT" (attribuierendes Demonstrativpronomen), "ART" (Artikel), "VVFIN" (finites Verb) und "APPRART" (Präposition mit Artikel). Durch den POS-Tagger ist es nun also möglich eine Suche im UMLS-Metathesaurus auf die Token Prozess, Leberzellverfettung und Fettleber zu beschränken. Zusätzlich zu Substantiven werden auch Token die mit einem Bindestrich abgeschlossen werden (Kohlehydrat- und Eiweißstoffwechsel) in die Menge der Begriffe die mit dem UMLS Metathesaurus analysiert werden aufgenommen. Hier verwendet der TreeTagger den Tag "TRUNC" (Kompositions-Erstglied).

[15] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`
[16] `http://code.google.com/p/tt4j/`



(a) Anzahl Token je Dokument nach Filterung (b) Anzahl an Token nach Anwendung von Stoppwortlisten und POS-Tagging

Abbildung 5.10: Vergleich der Effektivität von Filterungsmechanismen

5.3.3 Ergebnisse der Filterung

Anhand von fünf Testdokumenten wird nun die Auswirkungen das Stemming und POS-Tagging auf die Menge der verbleibenden Token ermittelt. Dies ist in Abbildung **??** dargestellt.

Es zeigt sich, dass der Prozess der Filterung erhebliche Auswirkungen auf die Menge der Token die mit dem UMLS Metathesaurus verglichen werden sollen hat. So reduziert sich die Menge an Token im Durchschnitt um 86%. Die grösste Auswirkung auf die Zahl der Token hat dabei das POS-Tagging was zu einer Reduktion um durschnittlich 74% führt. Bei einem Dokument, dass nach dem Tokenizing aus knapp 3600 Token besteht wird durch das POS-Tagging und Stemming eine Reduzierung auf knapp 450 Token erreicht. Diese 450 Token werden dann in einem nächsten Schritt zum Auffinden passender Konzepte im UMLS Metathesaurus verwendet. Weiterhin wird deutlich, dass eine Verwendung von Stopwortlisten keine Verringerung der Tokenanzahl bewirkt. Dies ist darauf zurückzuführen, dass das POS-Tagging bereits eine Filterung auf Substantive vorgenommen hat. Ein deutlicher Unterschied ist bei der Filterung der Inhalte zwischen der Verwendung von Stoppwortlisten und dem POS-Tagging zu erkennen, was in Abbildung 5.10 dargestellt ist.

Dabei wurde für das Testdokument 5 eine Filterung auf Basis von Stopwortlisten und POS-Tagging vorgenommen. Es ist zu erkennen, dass POS-Tagging die Menge an Token im Gegensatz zu Stoppwortlisten deutlich unterschreitet. Daher wird im Weiteren Verlauf auf die Nutzung von Stoppwortlisten verzichtet.

5.4 Merkmalanalyse

Nachdem in den vorangegangenen Abschnitten der Text in Token zerlegt und eine Filterung der Inhalte vorgenommen wurde bildet der folgende Abschnitt die zentrale Phase in der Verarbeitung natürlicher Sprache. Die bisherigen Abschnitte haben den Text so vorverarbeitet, dass nun auf Basis der Token eine tiefergehende Analyse der Inhalte vorgenommen werden kann. Bisher wurde jedoch noch keine Aussage über die Relevanz eines Token im Text gemacht. Wie in (Mihalcea und Tarau, 2004) oder (Medelyan et al., 2009) kann die Verwendung eines kontrollierten Vokabulars dabei unterstützend eingesetzt werden.

Ein vergleichsweise primitiver Ansatz die Relevanz eines Token zu ermitteln ist ein einfacher Abgleich mit dem UMLS Metathesaurus. Die zugrundeliegende Annahme lautet dabei, dass ein Token einen relevanten medizinischen Begriff darstellt, wenn ein passendes UMLS Konzept zu diesem Begriff gefunden werden konnte. Als zu analysierender Text wird der im Anhang 8.2 dargestellte Blogeintrag zum Thema "Fettleber" verwendet, der durch den Autor verschlagwortet ist als "Alkoholsucht", "Alkoholismus" und "Leber". Nach Anwendung der Filterungsmechanismen verbleiben 67 Token, für die ein Abgleich mit dem UMLS Metathesaurus durchgeführt wird. In Abbildung 5.11 ist ein Auszug der Ergebnisse eines Abgleichs mit dem UMLS Metathesaurus dargestellt.

Es wird deutlich, dass zu den analysierten Token eine Vielzahl an UMLS Konzepten identifiziert werden konnten. Insgesamt konnten zu 40 Token 99 UMLS Konzepte identifiziert werden[17]. Üblich für eine Verschlagwortung sind nach (Palshikar, 2007) 10 bis maximal 20 Schlagwörter. Weiterhin wird deutlich, dass zahlreiche Token nicht zwingend den thematischen Hintergrund des Textes wiedergeben (Abbau, Betroffene, Prozess etc.). Die größe der Knoten in dem dargestellten Graph ergibt sich aus der Zahl an idenitifizieren UMLS Konzepten zu dem verwendeten Token. Dies ist jedoch kein geeigneter Indikator für die Relvanz eines Token, da dadurch lediglich festgehalten werden kann, dass es sich um ein sehr allgemeines Konzept handelt.

Ein Merkmal der Relvanzbewertung, dass verbreitet im Bereich des Information Retrieval eingesetzt wird ist das sogenannte *TF/IDF* Maß (*Term Frequency - Inverse Document Frequency*). Das TF/IDF Maß setzt sich zusammen aus der Häufigkeit eines Token in dem zu analysierenden Dokument (Termfrequenz) und der Zahl aller Dokumente in denen der Token vorkommt (Dokumentfrequenz). Dabei gilt die Annahme, dass ein Term relevant ist wenn er in dem zu analysierenden Dokument häufig, aber über

[17] Token die lediglich als Zeichenfolge in einem UMLS Konzept auftreten werden ignoriert.

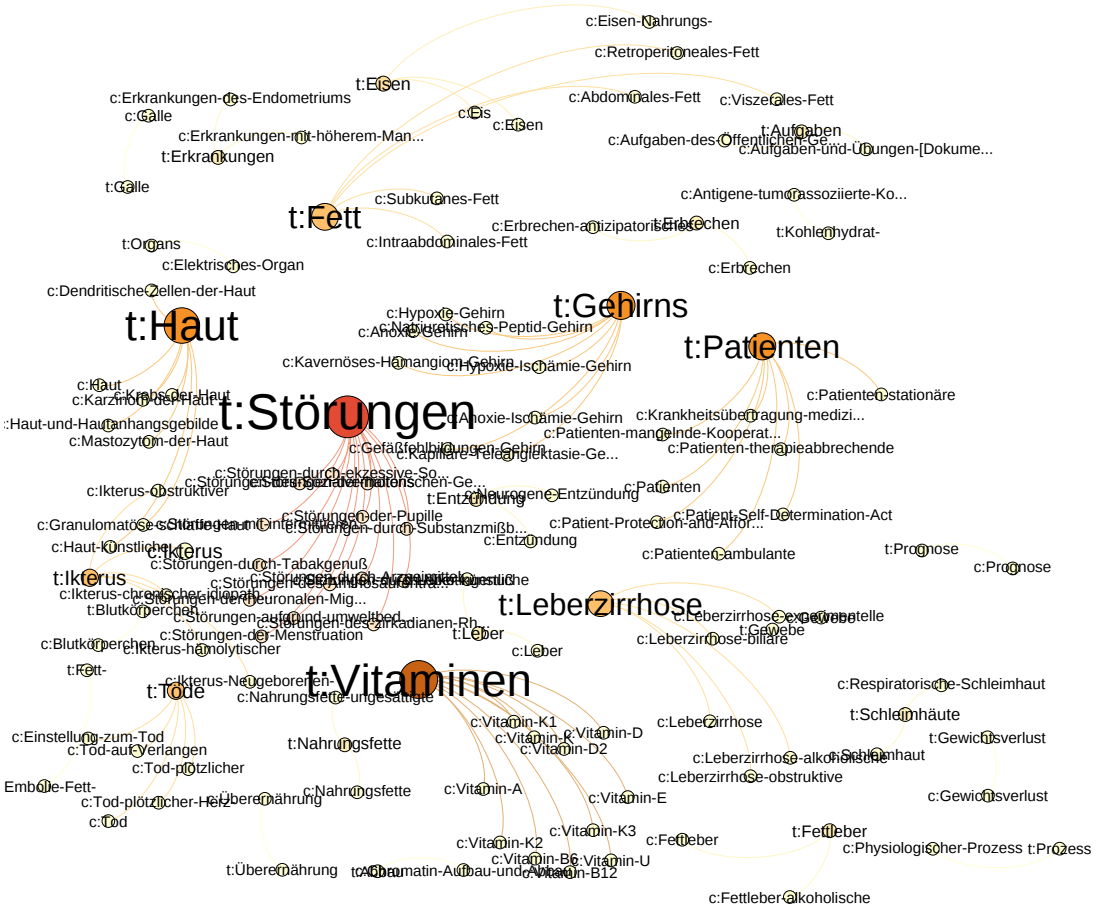


Abbildung 5.11: Gesamtmenge an Token (durch vorangestelltes "t" gekennzeichnet) und UMLS Konzepten (durch vorangestelltes "c" gekennzeichnet) zu dem Beispieltext

die Gesamtzahl der Dokumente seltener auftritt (Granitzer, 2006). Zur Berechnung der TF/IDF existieren zalhreiche Variationen, wobei die wohl am weitesten verbreitete in Gleichung 5.1 dargestellet ist.

$$w_{i,j} = tf_{i,j} * log\frac{N}{n_i} \tag{5.1}$$

Für einen tiefergehenden Einblick in das TF/IDF Maß sei hier auf (Robertson, 2004) verwiesen. Durch einen Perspektivenwechsel lässt sich der UMLS Metathesaurus aus Sicht eines Dokumentkorpus betrachten der eine Relevanzbewertung auf Basis der TF/IDF zulässt. Die Gesamtheit der Dokumente N ist somit definiert als die Menge aller Konzepte die zu einem Text ermittelt werden konnten. Eine angepasste Version

der TF/IDF kann daher lauten:

$$w_{i,j} = tf_{i,j} * log \frac{\text{Anzahl aller gefundenen Kozepte}}{\text{Anzahl gefundenen Konzepte zu } Token_i} \tag{5.2}$$

Auf diese Weise lässt sich die Menge gefundener Konzepte zu einem Token relativieren. Für den verwendeten Beispieltext ergibt eine Relevanzbewertung auf Basis der TF/IDF die in Abbildung 5.12 dargestellte Menge an Schlagworten. In der Abbildung wurden lediglich die sechs bestbewerteten Schlagwörter betrachtet. Im Gegensatz zu einem reinen Vergleich zwischen Token und dem Metathesaurus lässt sich erkennen, dass Begriffe die zu allgemein sind um zur Beschreibung des Textes verwendet werden zu können (bspw. "Störung" oder "Vitamin") relativiert werden konnten.

Abbildung 5.12: Ergebnis der Relevanzbewertung für die fünf bestbewerteten Token sowie derer UMLS Konzepte

Bisher wurde lediglich eine Aussage über die Relvanz von Token getroffen. Es wurde noch keine Aussage über die Relevanz der UMLS Konzepte getroffen. Dies hat zur Folge, dass sämtliche UMLS Konzepte, die zu einem Token gefunden wurden von gleicher Relevanz sind. So sind die Konzepte "Leberzirrhose-alkoholische" und "Leberzirrhose-experimentelle" von gleicher Relvanz. Der Folgende Ansatz beschäftigt sich mit der Relevanzbewertung der UMLS Konzepte.

5.4.1 Knotengrad

Ein Indikator für die Relevanz eines UMLS Konzepts ist der sogenannte Knotenrang innerhalb eines kontrollierten Vokabulars. So wird in *Maui* (Medelyan et al., 2009) sowie *KEA++*(Medelyan und Witten, 2006) die Annahme vertreten das ein Text der

ein bestimmtes Thema behandelt mit hoher Warscheinlichkeit eine Vielzahl von Token auweist die thematisch in Relation zueinander stehen. Auch der UMLS Metathesaurus weist Relationen zwischen Konzepten auf die verwendet werden können um eine thematische Zusammengehörigkeit zu identifizieren. Zu den unterstützten Relationstypen des Metathesaurus zählen Vater - Kind Beziehungen, Geschwister Beziehungen sowie Distanzbeziehungen (broader - narrower). In Abbildung 5.13 sind die Relationen zwischen den UMLS Konzepten der fünf bestbewerteten Token aus Abbildung 5.12 dargestellt.

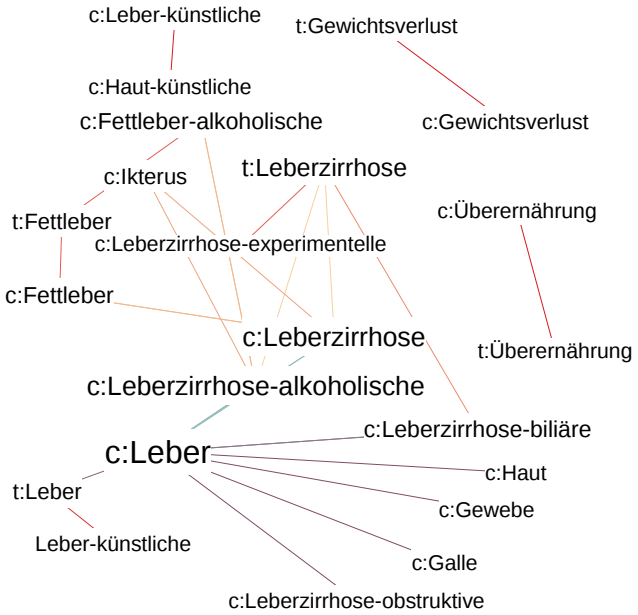


Abbildung 5.13: UMLS Konzept-Relationen zwischen den fünf bestbewerteten Token

Es lässt sich erkennen, dass die Konzepte "Gewichtsverlust" und "Überernährung" keinerlei Relationen zu anderen Konzepten aufweisen. Zwischen den Konzepten "Fettleber", "Leber" sowei "Leberzirrhose" hingegen konnten zahlreiche Relationen identifiziert werden.
Der Knotenrank wird errechnet indem jedem Konzept ein Relevanzwert zugewiesen wird, definiert als:

$$Konzept_{i,j} = \frac{\text{Anzahl Relationen im Text}}{\text{Anzahl Relationen im Methathesaurus}} \tag{5.3}$$

So wird zu jedem $Konzept_{i,j}$ eines $Token_i$ die Anzahl an Relationen im Metathesaurus ermittelt und dann unter Verwendung der Anzahl an Relationen im Text normalisiert. Auf diese Weise wird es nun möglich eine Relevanzbewertung nicht ausschließlich für Token sondern auch für UMLS Konzepte zu nutzen. Die Relevanzbewertung lässt sich

an dem Token "Leberzirrhose" verdeutlichen. In Tabelle 5.6 sind die UMLS Konzepte zu dem Token "Leberzirrhose" nach Relevanz sortiert dargestellt.

Rang	UMLS Konzept
1	Leberzirrhose
2	Leberzirrhose, alkoholische
3	Leberzirrhose, biliäre
4	Leberzirrhose, experimentelle

Tabelle 5.5: Relevanzbewertung von UMLS Konzepten zu dem Token Leberzirrhose

UMLS Konzepte, die in der Zeichenfolge vollständig mit dem gesuchten Token übereinstimmen werden bevorzugt behandelt. In diesem Fall wird das entsprechende Konzept unabhängig von seinem Relevanzwert grundsätzlich auf Rang eins dargestellt. Für das Konzept "Leberzirrhose, alkoholische" konnten insgesamt neun Relationen zu anderen im Text enthaltenen UMLS Konzepten identifiziert werden. Den geringsten Relevanzwert weist das UMLS Konzept "Leberzirrhose, experimentelle" mit drei Relationen zu anderen enthaltenen UMLS Konzepten auf.

5.4.2 Evaluation

Nachdem eine Relevanzbewertung von Token und UMLS Konzepten durchgeführt wurde folgt nun eine Messung der Effektivität der automatischen Verschlagwortung. Zu diesem Zweck werden fünf Zusammenfassungen (Abstract) aus medizinischer Fachliteratur zu den Themen Leber und Herz Chirurgie automatisch verschlagwortet. Die zusammengestellten Texte verfügen bereits über Schlagworte, was einen Vergleich mit der automatischen Verschlagwortung ermöglicht. Die Texte wurden so ausgewählt, dass sichergestellt ist, dass zuvor vergebene Schlagwörter auch innerhalb der Zusammenfassung erwähnt werden. Bei der Bewertung, ob ein automatisch vergebenes Schlagwort einem der bereits vergebenen Schlagwörter entspricht wird lediglich eine exakte Übereinstimmungen als positiver Wert interpretiert. Findet sich eine Abweichungen zwischen den Schlagwörtern wird dies als negativ bewertet (bspw. Blutung und Postoperative Blutung). Um morphologische Varianten in den Begriffen zu identifizieren werden beide Seiten mittels Stemming auf ihre Stammform transformiert.

Zur Messung der Effektivität der automatischen Verschlagwortung wird die Relevanzrate (*Precision*) und Vollständigkeitsrate (*Recall*) zu den fünf Testdokumenten berechnet. Bei der Relevanz- und Vollständigkeitsrate handelt es sich um Bewertungsmaße die häufig Anwendung im Bereich des Information Retrieval findet. Die Relevanzrate misst die Genauigkeit mit der ein System relevante Dokumente von nicht-relevanten Dokumen-

ten trennt (Gödert et al., 2012). Die Berechnung der Relevanzrate ist in Gleichung 5.4 dargestellt.

$$Relevanz = \frac{\text{Anzahl gefundener relevanter Dokumente}}{\text{Anzahl aller gefundenen Dokumente}} \tag{5.4}$$

Im Gegensatz zu der Relevanzrate misst die Vollständigkeitsrate das Verhältnis von gefundenen relevanten Dokumenten zu der Zahl an relevanten Dokumenten. Die Berechnung der Vollständigkeitsrate ist in Gleichung 5.5 dargestellt.

$$\text{Vollständigkeit} = \frac{\text{Anzahl gefundener relevanter Dokumente}}{\text{Anzahl aller relevanten Dokumente}} \tag{5.5}$$

In der Regel erfolgt eine Berechnung beider Bewertungsmaße, da jedes für sich genommen nur geringe Aussagekraft besitzt. So ist bspw. bei einer automatischen Verschlagwortung ein hohes Maß an Vollständigkeit nicht zwangsläufig ein Nachweis für die Effektiv. Konnten alle relevanten Begriffe in einem Text identifiziert werden führt dies zu einem hohen Maß an Vollständigkeit. Ist die Menge an Schlagwörtern die das System für relevant beachtet dabei besonders groß führt dies zu einer geringen Relevanzrate.

Die Ergebnisse der Vollständigkeits- und Genauigkeitsrate für die fünf Testdokumente sind in Tabelle 5.6 dargestellt. Die Berechnung wurde jeweils einmal für eine Obergrenze von fünf Token und für die Gesamtmenge an Token durchgeführt. Bei der Menge an UMLS Konzepten, die zu einem Token identifiziert werden konnten wurde keine Obergrenze festgelegt. Die Berechnung wurde für eine Ergebnissmenge von maximal fünf Token und für die Gesamtmenge an Token durchgeführt.

	Fünf bestbewertete Token		Gesamtmenge an Token	
Dok.	Genauigkeit	Vollständigkeit	Genauigkeit	Vollständigkeit
1	0,057	0,333	0,0645	0,666
2	0,2	0,333	0,0185	0,333
3	0,105	0,4	0,0074	0,4
4	0,214	0,6	0,0076	0,6
5	0,2	0,333	0,052	0,33

Tabelle 5.6: Vollständigkeits und Genauigkeitsrate für fünf Testdokumente

Im Durchschnitt umfassten die Testdokumente 3,8 Schlagwörter pro Dokument. Es zeigt sich, dass zu jedem Dokument mindestens ein Schlagwort identifiziert werden konnte welches auch von den Autoren gewählt wurde. Für Testdokument 4 sind manuell und automatisch vergebene Schlagwörter in Tabelle 5.7 dargestellt. Die Sortierung in Tabelle 5.7 entspricht der ermittelten Relevanzreihenfolge der automatisch vergebenen Schlagwörter.

Manuelle vergebene Schlagwörter	Automatisch ermittelte Schlagwörter
Aszites	**Aszites**
Drainage	**Drainage**
Nabelhernie	Morbidität
Leberzirrhose	**Leberzirrhose**
Hernienreparation	Patienten

Tabelle 5.7: Vergleich zwischen manuelle und automatisch vergebenen Schlagwörtern

Zu den fünf bestbewerteten Schlagworten für Testdokument 4 fanden sich insgesamt 14 UMLS Konzepte. Bei einigen Token wurde deutlich, dass verhältnismäßig viele UMLS Konzepte dazu existieren was sich in der Genauigkeitrate wiederspiegelt. Bei Betrachtung der Gesamtmenge an Token zeigt sich bspw., dass der Token "Gruppe" in 14 UMLS Konzepten identifiziert werden konnte[18]. Über die Gesamtmenge an Dokumenten konnten zu insgesamt 52 Token 168 UMLS Konzepte identifiziert werden.

Weiterhin konnte die Beobachtung gemacht werden, dass einige Token besonders häufig in den analysierten Dokumenten erwähnt werden. Beispielsweise der Token "Patient" wurde in sämtlichen analysierten Dokumenten erwähnt. Jedoch wurde er in keinem Dokument auch als Schlagwort verwendet. Innerhalb der automatischen Verschlagwortung hingegen zählte der Token "Patient" zweimal zu den fünf bestbewerteten Token. Eine genaue Beurteilung kann jedoch nur formuliert werden, wenn eine ausreichende Menge an Testdokumenten zur Verfügung steht und dies in Zusammenarbeit mit Domänenexperten erfolgt. Dennoch zeigen die ermittelten Werte, dass bereits zahlreiche positiv gewertete Schlagwörter identifiziert werden konnten.

[18] Es werden nur UMLS Konzepte betrachtet die den gesuchten Token als vollständige Zeichenkette beinhalten

6 Integration externer Daten

Der vorangegangene Abschnitt hat aufgezeigt wie mit Hilfe eines kontrollierten Vokabulars medizinisch relevante Begriffe in einem Text identifiziert werden können. Als Konsequenz werden diese nun zur Integration externer semantisch aufbereiteter Informationen verwendet. Hier wird der Umstand ausgenutzt, dass einige Wissensbasen im Web bei der Indizierung ihrer Inhalte ebenfalls eine konsistente Terminologie verwenden, beispielsweise PubMed und das kontrollierte Vokabular MeSH. Durch die Nutzung einer einheitlichen Terminologie auf Seiten der Anfrageformulierung und der zu integrierenden Wissensbasis kann die Präzision der Resultate erhöht werden. Neben PubMed bildet das *Linking Open Data* Projekt eine vielversprechende Informationsquelle für den Bereich der Medizin. Auch hier erfolgt eine Integration externer Informationen aus verschiedenen Wissensbasen wie *DBPedia*, *Diseasome* oder *Drugbank*. Im folgenden Abschnitt wird erläutert wie dies technisch realisiert wurde und welche Quellen bei der Integration verwendet werden.

6.1 PubMed

Eine der Kernkompetenzen die moderne Mediziner mitbringen müssen ist die Orientierung in großen bibliographischen Datenbeständen. Ein solcher bibliographischer Datenbestand ist das von dem *National Center for Biotechnology Information* (NCBI) entwickelte PubMed[1]. Mit zur Zeit[2] mehr als 21 Millionen Aufsatzzitaten aus über 5600 medizinischen Journalen hat sich PubMed zu einer zentralen Informationsquelle bei der Suche nach biomedizinischer Literatur entwickelt. Die Informationen die PubMed verwaltet sind jedoch lediglich als Sammlung von Metadaten zu den Artikeln zu verstehen, da PubMed keine Kopien der Artikel verwaltet. Zu den Metadaten zählen eine kurze Zusammenfassung des behandelten Themas (Abstract), sowie Angaben zu Autoren, Erscheinungsjahr und Journal. Neben den genannten Medatdaten bietet PubMed bei entsprechender Verfügbarkeit zusätzliche Informationen zu dem Artikel

[1] `http://www.ncbi.nlm.nih.gov/pubmed/`
[2] Stand 01.2012

an. Dies geschieht über das sogenannte *LinkOut* mit dessen Hilfe PubMed Inhalte von Autoren mit einem Verweis auf den Originalartikel oder auch die Webseite der Autoren versehen werden können. Bei den gesammelten Inhalten in PubMed handelt es sich um Metadaten zu sowohl kostenlosen als auch zahlungspflichtigen Artikeln. Letztere finden sich in *PubMed Central*[3] wieder, das zur Zeit 2,3 Mio. lizenzfreie Artikel umfasst.
Eine der Besonderheiten von PubMed ist die konsistente Indizierung der Inhalte auf Basis des kontrollierten Vokabulars MeSH. Im Gegensatz zu *Folksonomies*, bei der eine Verschlagwortung von Inhalten durch Nutzer ohne begriffliche Vorgaben vorgenommen wird geschieht die Indizierung in PubMed durch Experten unter Verwendung eines kontrollierten Vokabulars. Bei der Indizierung wird jeder Artikel mit einer Reihe von MeSH Termen angereichert die geeignet sind das behandelte Thema zu beschreiben (in der Regel 5 bis 15 MeSH Terme je Artikel). Bei der Anreicherung mit MeSH Termen wird zwischen *MeSH Headings*, *MeSH Subheadings* und *Major Topics* unterschieden. MeSH Headings dienen der Beschreibung des behandelten Themas wohingegen ein MeSH Subheading der näheren Beschreibung bestimmter Aspekte eines MeSH Headings dient. Dies lässt sich am Beispiel eines PubMed Artikels welcher in Abbildung 6.1 verkürzt dargestellt ist verdeutlichen.

Neben dem Titel und einer kurzen Zusammenfassung finden sich 10 MeSH Terme wieder die das behandelte Thema beschreiben. In dem Beispiel *Cardiovascular Diseases/chemically induced* ist das MeSH Heading gegeben durch *Cardiovascular Diseases* und das MeSH Subheading *chemically induced* durch einen "/" getrennt. Es ist zu erkennen, dass sich ein MeSH Subheading immer auf ein MeSH Heading bezieht mit dem Ziel dieses näher beschreiben zu können. Bei Major Topics handelt es sich um primäre MeSH Terme die zur Kennzeichnung von Termen verwendet werden welche das Thema am genauesten beschreiben. Diese sind in dem Beispieltext hervorgehoben.

Studien haben gezeigt,das ein Drittel aller PubMed Nutzer aus der allgemeinen Öffentlichkeit stammen. Die restlichen zwei Drittel verteilen sich auf Experten aus dem Gesundheitswesen und der Wissenschaft. Bei einer Analyse der Nutzungsstatistik wurde deutlich, dass medizinische Bibliothekare deutlich präzisere und vollständigere Ergebnisse bei der Suche nach PubMed Inhalten erzielen als medizinische Novizen. Dabei unterscheiden die Autoren zwischen informations- und navigationsorientierter Suche. Bei einer informationsorientierten Suche liegt der Fokus auf dem Auffinden von Informationen zu einem bestimmten Thema, wohingegen die navigationsorientierte Suche dem Auffinden von bestimmten Dokumenten dient (Herskovic et al., 2007). Die Nut-

[3] `http://www.ncbi.nlm.nih.gov/pmc/`

Title: Hormone therapy in perimenopausal and postmenopausal women: examining the evidence on cardiovascular disease risks.
Abstract: Women may live for 30 years or longer after menopause with cardiovascular disease as their highest mortality risk. Menopause may correspond to health alterations for women, yet the use of estrogen during and after this transition has been controversial for the past four decades. The evidence from recent scientific studies does not support the use of hormone therapy for the prevention or treatment of cardiovascular disease, which has resulted in its removal from national guideline recommendations.
Mesh Terms:
Aged
Cardiovascular Diseases/chemically induced
Estrogen Replacement Therapy/adverse effects
Evidence-Based Medicine
Female
Humans
Middle Aged
Perimenopause
Postmenopause
Risk Factors

Abbildung 6.1: PubMed Beispieleintrag Quelle: (NLM MeSH Indexing, 2008)

zung von PubMed macht schnell deutlich, dass es für Nutzer keine leichte Aufgabe ist genau die Informationen zu finden die benötigt werden. Dies ist zum einen auf die Menge an Inhalten die PubMed zur Verfügung stellt zurückzuführen. Es wird daher vermehrt dazu geraten die Suche durch Kombination mehrerer MeSH Terme zu verfeinern um spezifischere Suchergebnisse zu erhalten (Ebbert et al., 2003). Zum anderen zeigt sich das ein Abbilden von natürlicher Sprache auf Konzepte des kontrollierten Vokabulars für Nutzer keine leichte Aufgabe ist was dazu führt, dass bei der Suche häufig Begriffe verwendet werden die keinen gültigen MeSH Term darstellen (Gault et al., 2002). Zwar sind PubMed Inhalte konsistent indiziert doch wenn einem Nutzer die Terminologie eines Konzepts nicht bekannt ist erschwert dies das Auffinden von passenden Inhalten. Zum Abbilden von natürlicher Sprache auf MeSH Terme existiert eine Reihe von Möglichkeiten. Der *PubMed MeSH Browser* erlaubt beispielsweise eine Navigation durch die hierarchische Struktur um passende MeSH Terme zu finden. Weiterhin unterstützt PubMed Nutzer bei der Suche indem es passende MeSH Terme zu einem gegebenen Suchbegriff vorschlägt.

Eine der Fragen mit denen sich die vorliegende Arbeit auseinandersetzt ist wie Novizen bei einer informationsorientierten Suche unterstützt werden können und wie ein Abbilden natürlicher Sprache auf passende MeSH Terme vorgenommen werden kann.

6.1.1 Technische Umsetzung

Bei der Suche nach medizinischen Inhalten stellt die *National Library of Medicine* (NLM) ein Werkzeug namens *Entrez*[4] zur Verfügung welches Inhalte aus verschiedenen Datenquellen miteinander verknüpft und durchsuchbar macht . Entrez unterstützt die Suche in zahlreichen Datenbanken, darunter auch PubMed und PubMed Central. Verwaltet wird Entrez von dem *National Center for Biotechnology Information*[5] (NCBI) welches Teil der NLM ist. Im Fokus für die vorliegende Arbeit steht die Verwendung von PubMed sowie PubMed Central. Um eine Suche aus externen Anwendungen zu ermöglichen existiert eine Sammlung von acht Werkzeugen (E-Utilities), die als Schnittstelle zu Entrez und den verschiedenen Datenbanken dienen. Jede Anfrage an Entrez setzt sich zusammen aus einer statischen URL *http://eutils.ncbi.nlm.nih.gov/entrez/eutils/* , gefolgt von dem Namen des E-Utility und einer Reihe von Parametern, die der Spezifikation der Suchanfrage dienen. In einem ersten Schritt wird zunächst der deutschsprachige MeSH Term auf den entsprechenden englischen MeSH Term abgebildet. So wird der deutschsprachige MeSH Term "Fettleber" durch das englischsprachige Äquivalent "Fatty Liver" bei der Suche ersetzt. Der generelle Ablauf bei der Integration von PubMed Inhalten ist in Abbildung 6.2 dargestellt.

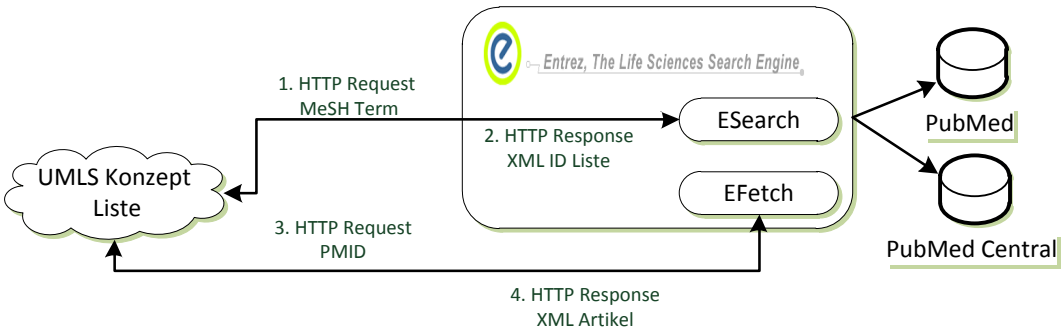


Abbildung 6.2: Prozessablauf bei der Integration von PubMed Inhalten

[4] `http://www.ncbi.nlm.nih.gov/sites/gquery`
[5] `http://www.ncbi.nlm.nih.gov/`

Zur Integration von PubMed Artikeln werden zwei Werkzeuge benötigt, *ESearch* und *EFetch* . Zunächst wird der englischsprachige MeSH Term via HTTP Request als Suchbegriff an das ESearch Utility übermittelt. Als Ergebnis der Suche liefert das ESearch Utility eine XML-basierte Liste von PMID's (PubMed Identifier) mit denen jeder PubMed Artikel eindeutig referenziert werden kann. Anhand der ermittelten PMID's werden dann unter Verwendung des EFetch Utility die entsprechenden Artikel abgerufen. Auch hier ist das Ergebnis eine auf XML basierende Liste. Das Ergebnis einer Suche nach dem MeSH Term "Fatty Liver" liefert über das ESearch Utility mehr als 46.000 PMID's. In Listing 6.1 ist das Ergebnis der EFetch Suche zu einem Artikel in verkürzter Form dargestellt. Der vollständige Eintrag findet sich in Anhang 8.1. Inhalte wie beispielsweise der Titel oder die Zusammenfassung eines Artikels wurden bereits in dem eingangs dargestellten Beispieltext erläutert. Bei den in Listing 6.1 dargestellten Informationen handelt es sich primär um Metadaten, die der näheren Beschreibung des Artikels dienen. Dazu zählt das Erscheinungsjahr sowie Erscheinungsform und Informationen zum Journal in dem der Artikel veröffentlicht wurde.

```
1  <PubmedArticle>
2          <PMID Version="1">22384276</PMID>
3          <DateCreated>
4              <Year>2012</Year>
5              <Month>03</Month>
6              <Day>02</Day>
7          </DateCreated>
8          <Article PubModel="Print-Electronic">
9              <Journal>
10                     <Volume>7</Volume>
11                     <Issue>2</Issue>
12                 <Title>PloS one</Title>
13                 <ISOAbbreviation>PLoS ONE</ISOAbbreviation>
14             </Journal>
15             <ArticleTitle>Preventing Phosphorylation of Sterol Regulatory
                   Element-Binding Protein 1a by MAP-Kinases Protects Mice
                  from Fatty Liver and Visceral Obesity.</ArticleTitle>
16         </Article>
17         <MedlineJournalInfo>
18             <NlmUniqueID>101285081</NlmUniqueID>
19         </MedlineJournalInfo>
20  </PubmedArticle>
```

Listing 6.1: PubMed Artikel zum MeSH Term "Fettleber" (verkürzt)

6.1.2 Filterung von PubMed Inhalten

Erste Gespräche mit potentiellen Nutzern haben deutlich gemacht, dass eine Integration von PubMed wünschenswert ist, jedoch zusätzlich eine Möglichkeit geschaffen werden muss die Menge an Inhalten zu Reduzieren. Eine Möglichkeit dies zu erreichen ist die Filterung auf bestimmte Journale. Jeder PubMed Artikel verfügt über eine Referenz auf die Quelle in der ein Artikel veröffentlicht wurde. Auf Basis einer *Journal-Whitelist* kann so eine Filterung vorgenommen werden. Neben einer Beschränkung auf bestimmte Journale erlaubt das EFetch Utility die Angabe eines Suchfeldes mit dem definiert werden kann zu welcher Kategorie der Suchbegriff gehört. Daher wird eine Beschränkung auf Artikel vorgenommen die den zu suchenden Begriff als primären MeSH Term indiziert haben. Als letzte Filteroption wird eine Beschränkung auf einen bestimmten Zeitraum verwendet. Hier wird als Initialwert ein Zeitraum von zwei Jahren festgelegt der jedoch individuell angepasst werden kann. Die Auswirkungen der hier dargestellten Filteroptionen werden deutlich, wenn man die Menge an gefundenen Artikeln zu dem MeSH Term "Fettleber" betrachtet. Die Ergebnisse einer Suche je nach Filtermecha-

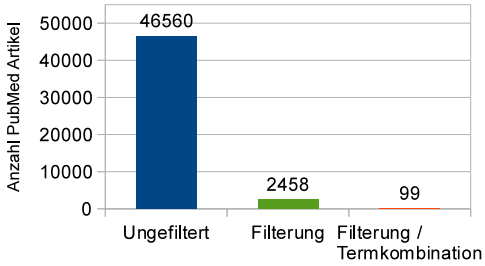


Abbildung 6.3: Filterung der Artikel in PubMed

nismus sind in Abbildung 6.3 dargestellt. Eine Suche ohne Filter liefert demnach mehr als 46.000 Artikel in denen der MeSH Term vorkommt wohingegen eine Suche unter Verwendung der Filteroptionen die Menge auf lediglich 2458 Artikel reduziert. Weiterhin kann die Verwendung von mehreren MeSH Termen bei der Suche die Präzision der Ergebnisse erhöhen. Verwendet man beispielsweise die MeSH Terme "Fettleber" und "Fettleber, alkoholische" gemeinsam als Suchbegriff reduziert sich auf diese Weise die Gesamtmenge auf knapp 100 Artikel. Auf diese Weise lassen sich komplexere Suchanfragen erstellen, die aus mehreren MeSH Headings, MeSH Subheadings und Major Topics bestehen. Beispielsweise kann dadurch eine Anfrage gebildet werden die nach "Fettleber" als Major Topic und "Fettleber, alkoholische" als MeSH Heading in Kombination mit dem MeSH Subheading "Arzneimitteltherapie" sucht.

6.2 Linked Open Data Project

Nicht zuletzt die Veröffentlichung der Linked Data Prinzipien (Berners-Lee, 2006) hat zu einem enormen Wachstum der Linked Data Cloud[6] geführt. So umfasst der Bereich der Lebenswissenschaften 41 Wissensbasen[7] mit Informationen zu Medikamenten, Krankheiten, klinischen Studien uvm., die mittels semantischer Technologien aufbereitet sind. Neben Wissensbasen der Lebenswissenschaften umfasst die Linked Data Cloud auch Informationen aus dem Bereich der nutzergenerierten Inhalte (User-generated Content). Hier zu nennen ist vor allem *DBPedia*, welches in RDF konvertierte Inhalte aus Wikipedia umfasst. Im Folgenden wird erläutert wie Informationen aus der Linked Data Cloud mittels Konzepten des UMLS Metathesaurus extrahiert werden können.

6.2.1 DBPedia

Wikipedia zählt ohne Zweifel zu den populärsten Webseiten im WWW. Nach[8] befindet sich Wikipedia auf Platz sechs der am häufigsten aufgerufenen Webseiten[9]. Heilman und Andere haben in (2011) Wikipedia als Informationsquelle für medizinische Zwecke evaluiert. Verschiedene Studien haben gezeigt, dass Wikipedia von 50-70% der praktizierenden Ärzte als Informationsquelle für medizinische Fragestellungen verwendet wird. Eine häufig geäußerte Kritik an Wikipedia ist die Gefahr der Ungenauigkeit, die sich aus der offenen Natur des Bearbeitungsprozesses ergibt. Im Wesentlichen zeigen die Studien jedoch, dass Wikipedia über eine grosse Zahl an Informationen aus dem Bereich der Medizin verfügt die nur geringe fachliche Fehler aufweisen.

Ein Problem bei der Nutzung von Wikipedia ist die Suche nach Inhalten. Hier offeriert Wikipedia lediglich eine Volltextsuche wodurch der Zugriff auf wertvolles Wissen deutlich eingeschränkt wird. Zwar ist Wissen in Wikipedia grundsätzlich vorhanden, jedoch kann es nicht durch Formulierung komplexer Suchanfragen abgerufen werden. Angenommen ein Nutzer möchte die Einwohnerzahl der Stadt Köln oder den Geburtsort von Angela Merkel erfahren. Dies sind zwar Informationen die sich in Wikipedia wiederfinden, jedoch nicht explizit als Suchanfrage formuliert werden können.

Das *DBPedia Project*[10] beschäftigt sich mit der Konvertierung von Wikipedia Inhalten in strukturiertes Wissen, welches mittels Semantic Web Technologien abgerufen werden

[6] `http://linkeddata.org/`
[7] (Stand 03/2012)
[8] `http://www.alexa.com/topsites`
[9] Stand 03/2012
[10] `http://dbpedia.org/About`

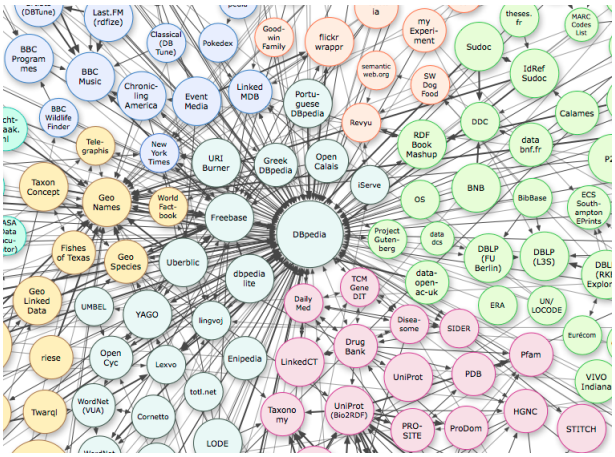


Abbildung 6.4: Ausschnitt der Linked Data Cloud - DPBedia (Richard Cyganiak, 2011)

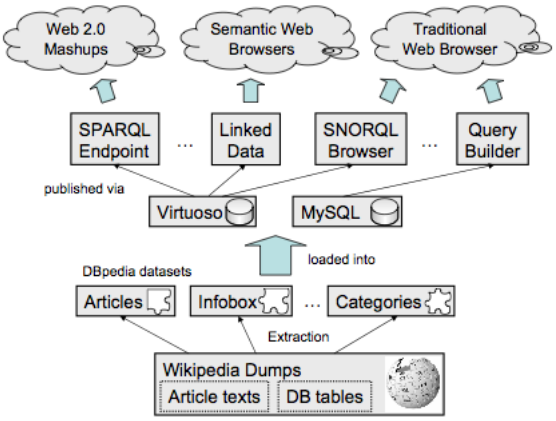


Abbildung 6.5: DPBedia Komponenten zur Informationsbeschaffung

kann. Um dies zu erreichen extrahiert DBPedia Information aus Wikipedia und bereitet diese in Form von RDF auf (Auer et al., 2007). Die damit gewonnene Wissensbasis wird nach den Linked Data Prinzipien aufbereitet und mit weiteren Wissensbasen verknüpft. Zur Zeit (Stand 03/2012) umfasst DBPedia circa 1 Milliarde RDF-Tripple zu mehr als 3,5 Millionen Ressourcen, was DBPedia zu der zentralen Wissensbases in der Linked Data Cloud macht. Dies wird in Abbildung 6.4 deutlich.

In Abbildung 6.5 sind die zentralen DBPedia Komponenten dargestellt. Für einen Zugriff auf Inhalte unter Verwendung von Semantic Web Technologien werden extrahierte Inhalte in einem Virtuoso Server[11] bereitgestellt. Über einen SPARQL Endpoint erlaubt

[11] `http://virtuoso.openlinksw.com/` Ein Datenbank Server der neben relationellen Datenbanken auch das Verwalten eines RDF Tripple Store erlaubt

DBPedia die Formulierung von SPARQL Abfragen zur Extraktion von Informationen aus RDF Dokumenten. Die SPARQL Abfrage zu dem UMLS Konzept *Fettleber* ist in Listing 6.2 dargestellt. In den Zeilen 5 und 6 erfolgt eine Filterung auf Deutsche und Englische Zusammenfassungen. In Zeile 8 ist dargestellt wie mittels SPARQL nach einer Zeichenfolge gesucht werden kann. Dazu wird die Systemfunktion *bif:contains* verwendet die in diesem Fall für das Prädikat *dbpedia2:meshname* eine Filterung auf Zeichenfolge *Fatty Liver* vornimmt. Optional werden *seeAlso* und *sameAs* Relationen sowie Weiterleitungen auf alternative Ressourcen ermittelt.

```
1  SELECT ?subject, ?mesh,?abstract_en,?abstract_de,?seeAlso,?sameAs,?depic
       ,?redirect WHERE {
2  ?subject  <http://www.w3.org/2000/01/rdf-schema#label> ?Literal .
3  ?subject <http://dbpedia.org/ontology/abstract> ?abstract_en.
4  ?subject <http://dbpedia.org/ontology/abstract> ?abstract_de.
5  ?subject dbpedia2:meshname ?mesh.
6  FILTER (LANG(?abstract_de) = 'de').
7  FILTER (LANG(?abstract_en) = 'en').
8  FILTER bif:contains(?mesh, "'Fatty Liver'")
9  optional{ ?subject owl:seeAlso ?seeAlso }
10 optional{ ?subject owl:sameAs ?sameAs }
11 optional{ ?subject <http://dbpedia.org/ontology/wikiPageRedirects> ?
       redirect}
12 }
13 }
```

Listing 6.2: Sparql Abfrage zu dem UMLS Konzept *Fettleber*

In Listing 6.3 ist die Extraktion des DBpedia Artikels zum dem UMLS Konzept *Fettleber* in verkürzter Form dargestellt. In Form von RDF Trippeln beinhaltet der dargestellte Auszug Informationen über die beschriebene Ressource, hier identifiziert über die URI *http://dbpedia.org/resource/Fatty_liver* . Wenn vorhanden weist DBpedia Ressourcen einen entsprechenden MeSH Term, sowie die zugehörige MeSH Tree Number zu. Innerhalb der DBpedia Ontologie sind diese als *meshName* und *meshNumber* modelliert. Wie an dem Triple *<foaf:name xml:lang="en">Fatty liver</foaf:name>* zu erkennen ist greift DBPedia bei der Beschreibung von Ressourcen auf bestehende, etablierte Ontologien wie FOAF[12] zurück. Da DBpedia Informationen aus mehrsprachigen Varianten von Wikipedia extrahiert beinhalten zahlreiche RDF-Triple ein Kürzel, das aussagt in welcher Sprache die Informationen vorliegen.

[12] `http://www.foaf-project.org/`

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:dbpprop="http://dbpedia.org/property/">
<rdf:Description rdf:about="http://dbpedia.org/resource/Fatty_liver">
<rdfs:label xml:lang="de">Fettleber</rdfs:label>
<foaf:name xml:lang="en">Fatty liver</foaf:name>
<dbpprop:meshname xml:lang="en">Fatty+Liver</dbpprop:meshname>
<dbpprop:meshnumber xml:lang="en">C06.552.241</dbpprop:meshnumber>
</rdf:RDF>
```

Listing 6.3: DBPedia Eintrag zum Thema "Fettleber" (verkürzt)

Um passende Inhalte in DBPedia zu identifizieren werden zuvor ermittelte UMLS Konzepte als Suchparamter an den SPARQL Endpoint gerichtet. MeSH Terme sind bei der Integration von Wikipedia zu bevorzugen. Es ist jedoch nicht garantiert, dass jedes Thema auch über einen MeSH Term verfügt. Verfügt eine Ressource in DBPedia nicht über das Prädikat *<dbpprop:meshname>* wird zusätzlich das Prädikat *<rdfs:label>* als Übereinstimmungskriterium verwendet.

Bei der Integration liegt der Fokus auf Informationen die es Nutzern ermöglicht sich einen oberflächlichen Eindruck zu dem gewählten UMLS Konzept zu verschaffen. Je nach Klassenzugehörigkeit verfügen Ressourcen über unterschiedliche Prädikate. So werden zur Beschreibung eines Organs andere Prädikate verwendet als bspw. bei einer Krankheit. Aus diesem Grund beschränkt sich die Integration auf allgemeine Inhalte wie eine kurze Zusammenfassung über die Ressource (Abstract), sowie alternative Bezeichnungen und wenn vorhanden multimediale Inhalte. Ein weiteres Ziel von Linked Data ist neben der Konvertierung von Daten in strukturiertes Wissen die Verknüpfung mit weiteren strukturierten Wissensbasen. Zu diesem Zweck verfügen DBPedia Ressourcen über sogenannte *sameAs* Prädikate mit denen eine Verknüpfung zu weiteren strukturierten Ressourcen oder auch unstrukturierten Webinhalten gewährleistet werden kann.

6.2.2 Linking Open Drug Data

Ein nicht unerheblicher Anteil an Informationen in der Linked Data Cloud beschäftigt sich mit Inhalten aus dem Bereich der Medizin. Innerhalb des sogenannten *Linking Open Drug Data* (LODD) Projekts wurden Datenquellen zu Medikamenten, klinischen

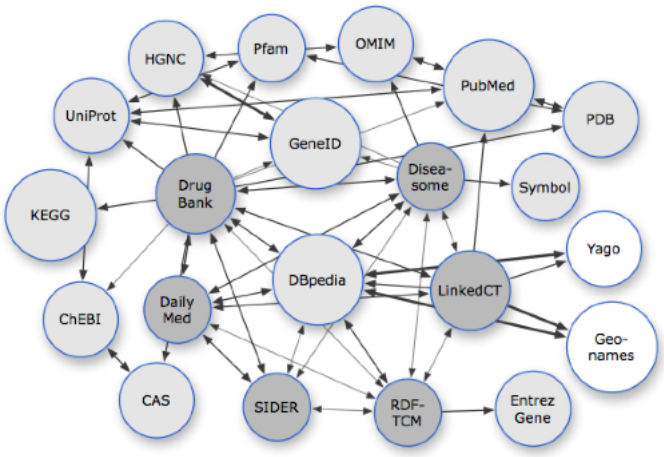


Abbildung 6.6: Teilausschnitt der Linked Data Cloud mit Fokus auf das Linking Open Drug Data Projekt

Studien, Krankheiten, traditioneller chinesischer Medizin und pharmazeutischen Unternehmen in die Linked Data Cloud integriert (Jentzsch et al., 2009). In Abbildung 6.6 sind die verschiedenen Quellen in der Linked Data Cloud dargestellt. Dunkelgraue Knoten repräsentieren Wissensbasen, die durch das LODD Projekt in die Linked Data Cloud integriert wurden. Insgesamt umfassen die Wissensbasen mehr als 8 Millionen Triple und knapp 400.000 Verknüpfungen zu externen Daten.

Einige der Wissenbasen werden nun auf Basis von UMLS Konzepten exemplarisch integriert.

Diseasome

Diseasome[13] umfasst ein Netzwerk von 4300 Genen und Genstörungen. Extrahiert wurden die Informationen aus der *Online Mendelian Inheritance in Man*[14] (OMIM) Datenbank. Da Diseasome durch eine Extraktion von OMIM Inhalten erzeugt wurde kann davon ausgegangen werden, dass bei der Bezeichnung von Ressourcen die OMIM Terminologie verwendet wurde. Dies führt dazu, dass eine konsistente Integration von Diseasome Inhalten nur garantiert werden kann, wenn eine Abbildung der MeSH Terminologie auf die OMIM Terminologie gegeben ist. Eine Möglichkeit dies zu erreichen ist die Verwendung des UMLS Metathesaurus zu dessen Quellen auch OMIM zählt. Auf Basis des Concept Unique Identifiers wird zu dem gegebenen UMLS Konzept die entsprechende OMIM Bezeichnung ermittelt und als Suchparameter für Diseasome ver-

[13] `http://www4.wiwiss.fu-berlin.de/diseasome/`

[14] `http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim` OMIM ist ein Kompendium über Gene und Genstörungen welches von dem *National Center for Biotechnology Information* (NCBI) vertrieben wird.

wendet. Dies lässt sich an dem Beispiel zu der Krankheit Alzheimer verdeutlichen. Das deutsprachige MeSH verwendet die Termbezeichnung "Alzheimer-Krankheit", was auf die von OMIM verwendete Bezeichnung "ALZHEIMER DISEASE" abgebildet werden kann.
Diseasome stellt einen Sparql Endpoint[15] zur Verfügung mit dessen Hilfe Anfragen an die Wissensbasis gestellt werden können. In einer ersten Phase erfolgt eine Ermittlung von Diseasome Ressourcen unter Verwendung der OMIM Terminologie. Für die Krankheit Alzheimer konnte eine Ressource ermittelt werden, die über den URI `http://www4.wiwiss.fu-berlin.de/diseasome/resource/diseases/74` identifiziert ist. Die entsprechende Sparql Abfrage ist in Listing 6.4 dargestellt. Über das Prädikat *rdfs:label* wird eine Suche nach dem Literal "Alzheimer Disease" durchgeführt. Zu der identifizierten Ressource werden dann alle zugehörigen Prädikate und Objekte ermittelt.

```
1 SELECT ?predicate ?object WHERE {
2  ?s rdfs:label "Alzheimer Disease".
3  ?s ?predicate ?object
4 }
```

Listing 6.4: Sparql Abfrage zu der Krankheit Alzheimer in Diseasome

Zu den Informationen die Diseasome zu Ressourcen bereitstellt zählen assozierte Gene, hierarchische Strukturinformationen, mögliche Arzneimittel sowie *sameAs* Relationen zu (externen) Ressourcen. Diseasome zeichnet sich durch eine reichhaltige Relationsbildung mit anderen semantisch aufbereiteten Wissensbasen aus. Beispielsweise verweisen Informationen zu Arzneimitteln mittels Angabe eines URI auf Ressourcen aus den Wissensbasen *Drugbank*[16] und *DailyMed*[17]. Der Prozess der Informationsextraktion über mehrere Wissensbasen hinweg ist in Abbildung 6.7 für die Ressource "Alzheimer Disease" vereinfacht dargestellt.

Die Integration von Informationen zu Arzneimitteln erfolgt über das Prädikat *diseasome:possibleDrug*. Zu der Alzheimer Krankheit umfasst Diseasome 110 Ressourcen die mittels eines URI auf die Wissensbasen Drugbank und Dailymed verweisen. In Abbildung 6.7 werden Informationen aus Drugbank und Dailymed exemplarisch zu den Ressourcen *dailymed:205* und *drugbank:DB00013* gewonnen. Für die Ressource *dailymed* wurden Informationen zu der Dosierungung des Medikaments und der Bezeichnung integriert. Neben Ressourcen die in externen Wissensbasen aufbereitet sind umfasst Diseasome auch Informationen über Gene und eine Kategorisierung von Krankheiten.

[15] `http://www4.wiwiss.fu-berlin.de/diseasome/snorql/`
[16] `http://www4.wiwiss.fu-berlin.de/drugbank/`
[17] `http://www4.wiwiss.fu-berlin.de/dailymed/`

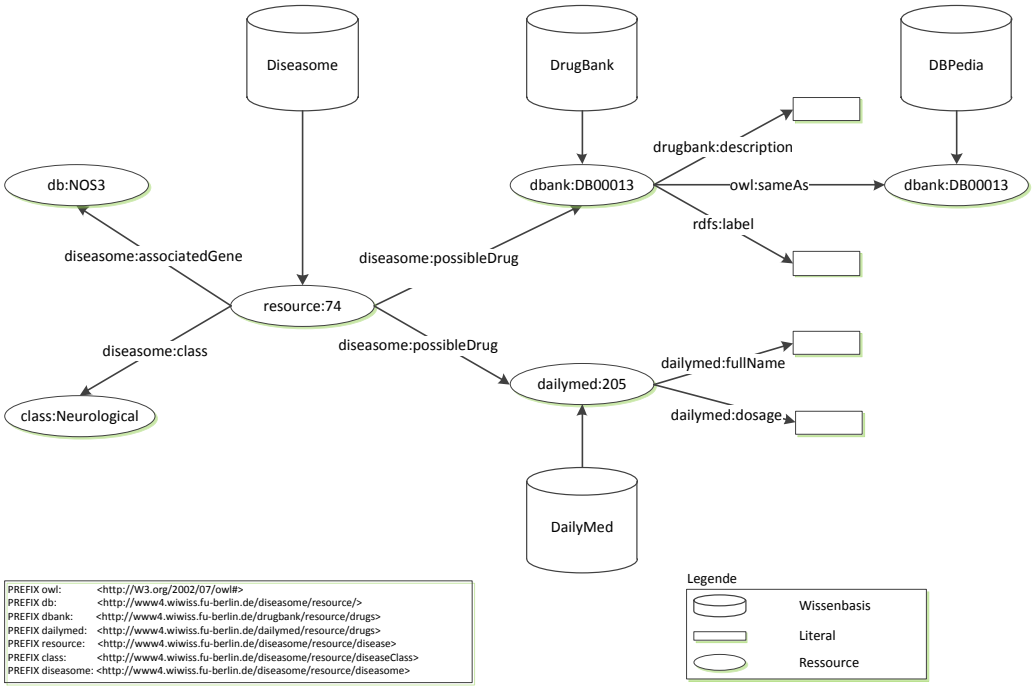


Abbildung 6.7: Integration semantischer Inhalte über mehrere Wissensbasen in Diseasome

Informationen über Gene werden durch das Prädikat *diseasome:associatedGene* ausgedrückt.

Für die Ressource *dbank:DB00013* wurde über die Prädikate *drugbank:description* und *drugbank:fullName* eine Beschreibung sowie eine Bezeichnung zu dem Medikament aus Drugbank ermittelt. Die dabei verwendete Sparql Abfrage ist in Listing 6.5 dargestellt. Die gewonnen Informationen sind in Tabelle 6.1 dargestellt. Es ist zu erkennen, dass auch Drugbank Relationen zwischen verteilten Ressourcen mittels eines *sameAs* Prädikats herstellt. In diesem Fall wird auf die DBPedia Ressource zu *Urokinase* verwiesen.

```
1 PREFIX drugbank: <http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs
      /DB00013>
2 SELECT ?description ?name ?same WHERE {
3   drugbank:DB00013 drugbank:description ?description.
4   drugbank:DB00013 rdfs:label ?name.
5   drugbank:DB00013 owl:sameAs ?same.
6 }
```

Listing 6.5: Sparql Abfrage zur Extraktion von Informationen in Drugbank

Prädikat	Objekt
drugbank:description	Low molecular weight form of human urokinase, that consists of an A chain of 2,000 daltons linked by a sulfhydryl bond to a B chain of 30,400 daltons. Recombinant urokinase plasminogen activator
drugbank:description	Urokinase
owl:sameAs	http://dbpedia.org/resource/Urokinase

Tabelle 6.1: Ergebnisse einer SPARQL Abfrage in Diseasome

Auch wenn das Linking Open Drug Data Projekt nur eine kleine Teilmenge der Linked Data Cloud ausmacht konnte gezeigt werden wie verschiedenen Wissensbasen mittels Relationen vernetzt sind. Eine intensive Betrachtung der verbleibenden Wissensbasen ist daher empfehlenswert.

7 Zusammenfassung

In der vorliegenden Arbeit wurde die Integration von externen semantischen Informationen auf Basis einer einheitlichen Terminologie behandelt. Um den Prozess der Abbildung natürlicher Sprache auf die zugrundeliegende Terminologie zu unterstützen wurde eine automatische Verschlagwortung textbasierter Inhalte vorgenommen. Mittels Methoden der Computerlinguistik konnte eine Vorverarbeitung textbasierter Inhalte durchgeführt werden. Bei der Identifikation medizinisch relevanter Begriffe in textbasierten Inhalten konnte ein domänenspezifisches Begriffssystem unterstützend eingesetzt werden. Durch Anwendung einer einheitlichen Terminologie auf Seiten des Anfragesystems und der Wissensbasis konnte der Prozess der Informationsbeschaffung optimiert werden. Im Folgenden werden die eingangs formulierten Fragestellungen rekapituliert.

7.1 Schlussfolgerung

Kann eine automatische Verschlagwortung textbasierter Inhalte auf Basis eines Begriffssystems optimiert werden?

Es zeigte sich, dass eine automatische Verschlagwortung durch die Verwendung eines domänenspezifischen Begriffsystems unterstützt werden kann. Insbesondere bei der Identifikation von Schlagwörtern, die keine Erwähnung im Text finden zeigt sich das Potential der Nutzung eines Begriffssystems. Weiterhin wurde ein verallgemeinertes Prozessmodell bei der automatischen Verschlagwortung vorgestellt. Die Phase der Vorverarbeitung erwies sich dabei als nichttrivialer Faktor für die Qualität der Verschlagwortung. So existieren Algorithmen, die ein optimiertes Tokenizing domänenspezifischer Inhalte unterstützen. Die Phasen des Stemming und POS-Tagging erwiesen sich als geeignetes Hilfsmittel zur Filterung textbasierter Inhalte. Verglichen mit einfachen Stoppwortlisten konnten deutliche Unterschiede hinsichtlich einer Reduzierung der Analysemenge erreicht werden. Die Relevanzbewertung automatisch vergebener Schlagwörter

erfolgt auf Basis der Präzision und Sensitivität. Wünschenswert ist hier eine Evaluation mit potentiellen Nutzern. Nur auf diese Weise lassen sich Aussagen über die Qualität von Schlagwörtern treffen die nicht explizit im Text oder der Menge vordefinierter Schlagwörter auftreten. Denkbar ist die Durchführung von Interviews mit potentiellen Nutzern, sowie ein Vergleich von automatisch vergebeben Schlagwörtern zu einer manuellen Verschlagwortung. Weiterhin ist ein Vergleich zwischen Algorithmen des maschinellen Lernens, der Empfehlungsdienste und graphbasierter Ansätze zur Extraktion textbasierter Inhalte ein interessanter Aspekt, der Raum für Folgearbeiten bietet. Bei Ansätzen des maschinellen Lernens und der Empfehlungsdienste ist jedoch das Vorhandensein entsprechender Trainingsdaten respektive Referenzbeiträge vorrausgesetzt. Als vielversprechend erweist sich der graphbasierte Ansatz TextRank (Mihalcea und Tarau, 2004), der eine Verschlagwortung auf Basis des PageRank Algorithmus (Brin und Page, 1998) vornimmt. Mihalcea und Tarau erzeugen zur Extraktion von Schlagwörtern einen ungerichteten Graphen unter Betrachtung von Kookurenzen. Eine interessante Fragestellung ist hier wie ein (un)gerichteter Graph auf Basis des UMLS Metathesaurus erzeugt werden kann.

Inwieweit kann die Nutzung einer einheitlichen Terminologie zwischen Anfragesystem und Wissensbasis den Prozess der Informationsbeschaffung unterstützen?

Bei der Integration externer semantischer Informationen zeigte sich, dass die Verwendung einer einheitlichen Terminologie erhebliche Vorteile bietet. Insbesondere zur Reduzierung von Unsicherheit bei der Informationsbeschaffung zeigten sich Vorteile gegenüber klassischen Informationsbeschaffungssystemen. Die Menge an Wissen, dass durch PubMed und die Linked Data Cloud zur Verfügung gestellt wird macht deutlich, dass eine Orientierung in solch großen Datenbeständen nicht ohne Weiteres möglich ist. Es wurde zudem ersichtlich, dass die Menge an Wissen, dass sowohl in PubMed als auch in der Linked Data Cloud zur Verfügung steht nicht ohne unterstützende Systeme sinnvoll verarbeitet werden kann. Die Ausdrucksstärke heutiger Semantic Web Technologien hat dazu geführt, dass zahlreiche Wissensbasen mittels semantischer Relationen vernetzt sind. Dieser Umstand wurde ausgenutzt um Informationen aus verschiedenen Wissensbasen zu aggregieren.

Durch die Verwendung der MeSH Terminolgie konnte eine Integration von PubMed Inhalten erreicht werden. Insbesondere der Detaillierungsgrad mit dem eine Suchanfrage formuliert werden kann hat zu einer deutlichen Reduzierung der Informationsmenge

gegenüber der Suche mittels einfacher Termbezeichner geführt. Dies erlaubt eine vereinfachte Orientierung in der Informationsmenge. Neben Informationen aus der Linked Data Cloud und PubMed konnten aus MeSH stammende Begriffsdefinitionen integriert werden. Diese erlauben das Schaffen eines grundlegenden Überblicks über behandelte Konzepte. Bei der Integration externer semantischer Inhalte bedarf es einer genauen Interpretation der Informationen. Die ausgewählten Wissensbasen liefern zweifelsohne interessante Informationen zu gegebenen medizinischen Konzepten. Es ist jedoch nicht gewährleistet, dass diese auch für jeden der zahlreichen medizinischen Teilbereiche gleich relevant sind. Die integrierten Informationen bilden zudem lediglich eine Teilmenge des verfügbaren Wissens der Linked Data Cloud. Hier wäre es sinnvoll Domänenexperten einen Zugriff auf die Linked Data Cloud zu ermöglichen um Wünsche oder Anregungen äussern zu können.

Welche medizinischen Begriffssysteme existieren und wie eignen sich diese als zugrungeliegendes Vokabular für die automatische Verschlagwortung und Integration semantischer Inhalte?

Eine Untersuchung verfügbarer medizinischer Begriffssysteme hat gezeigt das Standardisierungsbemühungen bestehen die sich mit der Etablierung einer einheitlichen medizinischen Terminologie befassen. Insbesonder der UMLS erwies sich als eine umfassende Quelle für medizinische Konzepte. Durch eine reichhaltige Modellierung von Wissen aus zahlreichen Quellen ist der UMLS Metathesaurus in der Lage zahlreiche Relationen zwischen Konzepten zu identifizieren sowie unterschiedliche Schreibweisen zu einem Konzept zu berücksichtigen. Das MeSH hat sich bei der Integration von PubMed Inhalten als geeignet erwiesen. Die in dieser Arbeit integrierten Quellen des UMLS haben gezeigt, dass diese in einer ausreichenden Detailtiefe modelliert sind um allgemeine textbasierte Inhalte auf Konzepte abzubilden. Die Tiefe in der Konzepte des UMLS Metathesaurus beschrieben sind bedarf einer Bewertung durch Domänenexperten. Hier gilt es insbesondere zu betrachten welche der zahlreichen Quellen des UMLS einen Mehrwert darstellen. Nur auf diese Weise lässt sich auch eine qualitative Aussage über den Detaillierungsgrad der Konzepte treffen.

Abschließend lässt sich zusammenfassen, dass die Integration von externen Informationen mittels einer einheitlichen Terminologie erfolgreich umgesetzt werden konnte. Die Menge an Informationen, die durch das Linking Open Data Projekt für den Bereich der Medizin zur Verfügung gestellt wird erwies sich dabei als umfangreich und

brauchbar. Durch die Verwendung von semantischen Relationen zwischen Ressourcen verteilter Wissensbasen konnten neue Informationen von externen Quellen erfolgreich integriert werden. Insbesondere das Linking Open Drug Data Projekt hat aufgezeigt wie dies technisch umgesetzt werden kann. Die Möglichkeiten die sich dadurch ergeben sind vielversprechend. Beispielsweise die Verknüpfung von Krankheiten mit entsprechenden Medikamenten und klinischen Studien stellt einen interessanten Aspekt dar. Hinsichtlich der Informationsbeschaffung im Bereich von PubMed konnte eine Optimierung verglichen mit einfachen Suchanfragen erreicht werden. Man kann davon ausgehen, dass dadurch eine wesentliche Systemunterstüzung für medizinisches Fachpersonal geschaffen werden konnte. Die automatische Verschlagwortung hat aufgezeigt, dass sich medizinisch relevante Begriffe in Text erfolgreich identifizieren lassen. Auf diese Weise wird das Abbilden natürlicher Sprache auf eine standardisierte Terminologie gewährleistet. Dies führt zu einem Mehrwert für Mediziner, da es den mühseligen Prozess des Identifizierens einer passenden Terminologie automatisiert.

Abbildungsverzeichnis

Literaturverzeichnis

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, und Zachary Ives. Dbpedia: A nucleus for a web of open data. *The Semantic Web*, 4825 (Springer):722–735, 2007.

Ricardo Baeza-Yates und Berthier Ribeiro-Neto. *Modern Information Retrieval*, volume 463. Addison Wesley, 1999.

Neil Barrett und Jens Weber-Jahnke. Building a biomedical tokenizer using the token lattice design pattern and the adapted viterbi algorithm. *BMC Bioinformatics 2011*, 12(June), 2009.

Christoph Beierle und Gabriele Kern-Isberner. Maschinelles Lernen. In *Methoden wissensbasierter Systeme*, pages 97–154. Vieweg, 2006. ISBN 978-3-8348-9116-7.

Tim Berners-Lee. Linked data - desing issues, 2006. URL `http://www.w3.org/DesignIssues/LinkedData.html`. Abgerufen 03/2012.

Tim Berners-Lee, James Hendler, und Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001.

Christian Bizer, Tom Heath, und Tim Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

Christian Bizer, Anja Jentzsch, und Richard Cyganiak. State of the lod cloud, 2011. URL `http://www4.wiwiss.fu-berlin.de/lodcloud/state/`. Abgerufen 02/2012.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270, 2004.

S Brin und L Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.

Claire Cardie. Empirical methods in information extraction. *AI Magazine*, 18(4):65–79, 1997.

Kai-Uwe Carstensen, Susanne J. Jekat, und Ralf Klabunde. Computerlinguistik – Was ist das? In *Computerlinguistik und Sprachtechnologie*, pages 1–25. Spektrum Akademischer Verlag, 2010. ISBN 978-3-8274-2224-8.

Andreas Dorm. Gopubmed: Ontology-based literature search for the life sciences. 2008. URL `http://nbn-resolving.de/urn:nbn:de:bsz:14-ds-1232454035091-47450`.

Jon O Ebbert, Denise M Dupras, und Patricia J Erwin. Searching the medical literature using pubmed: a tutorial. *Mayo Clinic proceedings Mayo Clinic*, 78(1):87–91, 2003.

Wolfgang Ertel. Maschinelles Lernen und Data Mining. In *Grundkurs Künstliche Intelligenz*, pages 179–240. Vieweg+Teubner, 2008. ISBN 978-3-8348-9441-0.

Usama Fayyad, Gregory Piatetsky-Shapiro, und Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54, 1996.

The Apache Software Foundation. About the apache incubator, a. URL `http://incubator.apache.org/incubation/Incubation_Policy.html`. zuletzt abgerufen 01.2012.

The Apache Software Foundation. Apache opennlp, b. URL `http://incubator.apache.org/opennlp/`. Abgerufen 01.2012.

Norbert Fuhr. Ziele und Aufgaben der Fachgruppe Information Retrieval, 1996. URL `http://www.uni-hildesheim.de/fgir/index.php?option=com_content&task=view&id=14&Itemid=41`. Abgerufen 03/2012.

Deutsches Institut für Medizinische Dokumentation und Information. Mesh - medical subject headings. URL `http://www.dimdi.de/static/de/klassi/mesh_umls/mesh/`. Abgerufen 02.2012.

Lora V Gault, Mary Shultz, und Kathy J Davies. Variations in medical subject headings (mesh) mapping: from the natural language of patron terms to the controlled vocabulary of mapped lists. *Journal of the Medical Library Association*, 90(2):173–180, 2002.

Barbara Geyer-Hayden. *Wissensmodellierung im Semantic Web*, pages 127–146. Springer, 2009.

Transinsight GmbH. Gopubmed, 2012. URL `http://www.gopubmed.org/web/gopubmed/`. Abgerufen 02/2012.

Scott A. Golder und Bernardo A. Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2):198–208, 2006.

Dr. Thomas Gottron. Vorlesungsskript - Information Retrieval. Johannes Gutenberg Universität Mainz - Institut für Informatik, 2010.

Michael Granitzer. Statistische Verfahren der Textanalyse. In Tassilo Pellegrini und Andreas Blumauer, editors, *Semantic Web*, X.media.press, pages 437–451. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-29325-5.

Gregory Grefenstette und Tapanainen Pasi. What is a word, what is a sentence? problems of tokenization. *Proceedings of 3rd Conference on Computational Lexicography and Text Research*, 3(August):79–87, 1994.

Thomas R Gruber. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(5-6):907–928, 1995.

Winfried Gödert, Klaus Lepsky, Matthias Nagelschmidt, Winfried Gödert, Klaus Lepsky, und Matthias Nagelschmidt. Retrievalexperimente. In *Informationserschließung und Automatisches Indexieren*, X.media.press, pages 327–346. Springer Berlin Heidelberg, 2012. ISBN 978-3-642-23513-9.

Ying He und Mehmet Kayaalp. A comparison of 13 tokenizers on medline. *U.S. National Library of Medicine*, 2006.

Marti A Hearst. *Untangling text data mining*, pages 3–10. Number Hearst. Association for Computational Linguistics, 1999.

James M Heilman, Eckhard Kemmann, Michael Bonert, Anwesh Chatterjee, Brent Ragar, Graham M Beards, David J Iberri, Matthew Harvey, Brendan Thomas, Wouter Stomp, und et al. Wikipedia: a key tool for global public health promotion. *Journal of Medical Internet Research*, 13(1):e14, 2011.

Andreas Henrich. Lehrbuch Information Retrieval 1. *Otto-Friedrich-Universität Bamberg, Lehrstuhl für Medieninformatik*, 11(5):2001–2008, 2008.

Jorge R Herskovic, Len Y Tanaka, William Hersh, und Elmer V Bernstam. A day in the life of pubmed: analysis of a typical day's query log. *Journal of the American Medical Informatics Association*, 14(2):212–220, 2007.

Hajo Hippner und René Rentzmann. Text Mining. *Informatik-Spektrum*, 29:287–290, 2006. ISSN 0170-6012. 10.1007/s00287-006-0091-y.

Pascal Hitzler, Markus Krötzsch, Sebastian Rudolph, und York Sure. Die Idee des Semantic Web. In *Semantic Web*, eXamen.press. Springer Berlin Heidelberg, 2008. ISBN 978-3-540-33994-6.

Andreas Holzinger, Regina Geierhofer, und Maximilian Errath. Semantische Informationsextraktion in medizinischen Informationssystemen. *Informatik-Spektrum*, 30: 69–78, 2007. ISSN 0170-6012.

Leonie IJzereef, Jaap Kamps, und Maarten De Rijke. Biomedical retrieval: How can a thesaurus help? *On the Move to Meaningful Internet Systems 2005 CoopIS DOA and ODBASE OTM Confederated International Conferences CoopIS DOA and ODBASE 2005 Proceedings Part II Agia Napa Cyprus*, 3761:1432–1448, 2005.

Rezarta Islamaj Dogan, G Craig Murray, Aurélie Névéol, und Zhiyong Lu. Understanding pubmed® user search behavior through log analysis. *Database the journal of biological databases and curation*, 2009(0):18.

Anja Jentzsch, Jun Zhao, Oktie Hassanzadeh, Kei-Hoi Cheung, Matthias Samwald, und Bo Andersson. Linking open drug data. In *Proceedings of Linking Open Data Triplification Challenge at the I-Semantics 2009*, 09 2009.

Jing Jiang und ChengXiang Zhai. An empirical study of tokenization strategies for biomedical information retrieval. *Information Retrieval*, 10:341–363, 2007. ISSN 1386-4564. 10.1007/s10791-007-9027-7.

Tomanel K. Documentation for jtbd 1.6. 2007. URL `https://julielab.de/Resources/Software/NLP+Tools/Download/Stand_alone+Tools.html`. Dokumentation als Teil der Installationsdatein - zuletzt abgerufen 01.2012.

Wolfgang Kienreich und Markus Strohmaier. Wissensmodellierung — Basis für die Anwendung semantischer Technologien. In Tassilo Pellegrini und Andreas Blumauer, editors, *Semantic Web*, X.media.press, pages 359–371. Springer Berlin Heidelberg, 2006. ISBN 978-3-540-29325-5.

Stefan Klatt und Bernd Bohnet. You don't have to think twice if you carefully tokenize. *Lecture Notes in Computer Science*, 3248:299–309, 2005.

Jena University Language & Information Engineering Lab. Nlp toolsuite. URL `https://julielab.de/Resources/Software/NLP_Tools.html`. zuletzt abgerufen 01.2012.

Richard Lenz, Mario Beyer, Christian Meiler, Stefan Jablonski, und Klaus A. Kuhn. Informationsintegration in Gesundheitsversorgungsnetzen. *Informatik-Spektrum*, 28: 105–119, 2005. ISSN 0170-6012.

Y Matsuo und M Ishizuka. Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, 13(1):157–169, 2004.

Deborah L McGuinness. *Ontologies come of age*, page 171–192. MIT Press, 2002.

Olena Medelyan und Ian H Witten. *Thesaurus based automatic keyphrase indexing*, page 296. ACM Press, 2006.

Olena Medelyan, Eibe Frank, und Ian H Witten. Human-competitive tagging using automatic keyphrase extraction. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 3 EMNLP 09*, 3(August):1318, 2009.

U S National Library Of Medicine. Umls reference manual, 2011. URL `http://www.ncbi.nlm.nih.gov/books/NBK9684/`. Abgerufen 12/2011.

Alexander Mehler und Christian Wolff. Perspektiven und Positionen des Text Mining. *GLDVJournal for Computational Linguistics and Language Technology*, 20(1):1–18, 2005.

Ryszard S. Y. Kodratoff Michalski. *Research in machine learning; recent progress, classification of methods, and future directions*, volume 3. Morgan Kaufmann, 1990.

Rada Mihalcea und Paul Tarau. Textrank: Bringing order into texts. *Conference on Empirical Methods in Natural Language Processing*, pages 404–411, 2004.

Dr.Raimund Mildner. Surgerytube Web 2.0 Technologien in der Qualifizierung von Chirurgen. *Projektantrag SurgeryTube*, 2009.

Gilad Mishne. Autotag: a collaborative approach to automated tag assignment for weblog posts. *Computer*, pages 953–954, 2006.

Vassil Momtchev, Deyan Peychev, Todor Primov, und Georgi Georgiev. Expanding the pathway and interaction knowledge in linked life data. *ontotextcom*, 2009.

National Library of Medicine. Unified medical language system basics. 2008. URL `http://www.nlm.nih.gov/research/umls/new_users/online_learning/OVR_001.htm`. Abgerufen 03/2012.

National Library of Medicine. Unified medical language system statistics - 2011ab release, 2011a. URL `http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html`. Abgerufen 03/2012.

U.S. National Library of Medicine. Fact sheet medical subject heading. a. URL `http://www.nlm.nih.gov/pubs/factsheets/mesh.html`. Abgerufen 02/2012.

U.S. National Library of Medicine. Statistics - 2011ab release. b. URL `http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/statistics.html`. Abgerufen 02/2012.

U.S. National Library of Medicine. Indexing with mesh vocabulary, 2011b. URL `http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/015_030.html`. Abgerufen 03/2012.

Tim O'Reilly. *What Is Web 2.0*, volume 30. San Francisco:, 2005. URL `http://oreilly.com/web2/archive/what-is-web-20.html`.

Girish Palshikar. Keyword extraction from a single document using centrality measures. In Ashish Ghosh, Rajat De, und Sankar Pal, editors, *Pattern Recognition and Machine Intelligence*, volume 4815 of *Lecture Notes in Computer Science*, pages 503–510. Springer Berlin / Heidelberg, 2007. ISBN 978-3-540-77045-9.

Claude Pasquier. Single document keyphrase extraction using sentence clustering and latent dirichlet allocation. *Computational Linguistics*, (July):154–157, 2010.

M F Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

Klaus Reichenberger. Erschließung von Dokumenten. In *Kompendium semantische Netze*, X.media.press, pages 125–141. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-04315-4.

Thomas Reinartz. Focusing solutions for data mining. volume 1623 of *Lecture Notes in Computer Science*, pages 69–69. Springer Berlin / Heidelberg, 1999. ISBN 978-3-540-66429-1.

Anja Jentzsch Richard Cyganiak. Linking open data cloud diagram. 2011. URL `http://lod-cloud.net/`.

Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation*, 60(5):503–520, 2004.

Helmut Schmid. *Improvements in Part-of-Speech Tagging with an Application to German*, volume 11, pages 47–50. Kluwer Academic Publishers, 1995.

L. Smith, T. Rindflesch, und W. J. Wilbur. Medpost: a part-of-speech tagger for biomedical text. *Bioinformatics*, 20(14):2320–2321, 2004. doi: 10.1093/bioinformatics/bth227. zuletzt abgerufen 01.2012.

Sanjay C Sood, Sara H Owsley, Kristian J Hammond, und Larry Birnbaum. Tagassist: Automatic tag suggestion for blog posts. *International Conference on Weblogs and Social Media*, 2007.

Cord Spreckelsen und Klaus Spitzer. Medizinische Wissensverarbeitung — Anwendungsszenarien. In *Wissensbasen und Expertensysteme in der Medizin*, pages 9–25. Vieweg+Teubner, 2009. ISBN 978-3-8348-9294-2.

Katrin Tomanek, Joachim Wermter, und Udo Hahn. A reappraisal of sentence and token splitting for life sciences documents. *Studies In Health Technology And Informatics*, 129(Pt 1):524–528, 2007.

Peter D. Turney. Learning to extract keyphrases from text. *National Research Council, Institute for Information Technology*, December 1999.

Jakob Voss. Tagging, folksonomy & co - renaissance of manual indexing? *CoRR*, abs/cs/0701072, 2007.

Ian H Witten, Gordon W Paynter, Eibe Frank, Carl Gutwin, und Craig G Nevill-Manning. Kea: Practical automatic keyphrase extraction. *Proceedings of the fourth ACM conference on Digital libraries*, Berkeley,:9, 1999.

Roman Yangarber, Ralph Grishman, Pasi Tapanainen, und Silja Huttunen. *Automatic Acquisition of Domain Knowledge for Information Extraction*, volume 2, page 940. Association for Computational Linguistics, 2000.

Anhang

8.1 Alzheimer Disease MeSH 2011

Attribut	Wert
Main Heading	Alzheimer Disease
Tree Number	C10.228.140.380.100
Tree Number	C10.574.945.249
Tree Number	F03.087.400.100
Scope Note	A degenerative disease of the BRAIN characterized by the insidious onset of DEMENTIA. Impairment of MEMORY, judgment, attention span, and problem solving skills are followed by severe APRAXIAS and a global loss of cognitive abilities. The condition primarily occurs after age 60, and is marked pathologically by severe cortical atrophy and the triad of SENILE PLAQUES; NEUROFIBRILLARY TANGLES; and NEUROPIL THREADS. (From Adams et al., Principles of Neurology, 6th ed, pp1049-57)
Entry Term	Acute Confusional Senile Dementia
Entry Term	Alzheimer Disease, Early Onset
Entry Term	Alzheimer Disease, Late Onset
Entry Term	Alzheimer Type Senile Dementia
Entry Term	Alzheimer's Disease
Entry Term	Alzheimer's Disease, Focal Onset
Entry Term	Dementia, Alzheimer Type
Entry Term	Dementia, Presenile
Entry Term	Dementia, Primary Senile Degenerative
Entry Term	Dementia, Senile
Entry Term	Early Onset Alzheimer Disease
Entry Term	Focal Onset Alzheimer's Disease
Entry Term	Late Onset Alzheimer Disease
Entry Term	Presenile Alzheimer Dementia
Entry Term	Primary Senile Degenerative Dementia
Entry Term	Senile Dementia, Acute Confusional
Entry Term	Senile Dementia, Alzheimer Type
See Also	Amyloid beta-Peptides
See Also	Amyloid beta-Protein Precursor
See Also	Aphasia, Primary Progressive
See Also	Cerebral Amyloid Angiopathy
See Also	Kluver-Bucy Syndrome
See Also	Neurofilament Proteins
See Also	tau Proteins
Allowable Qualifiers	BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH UR US VE VI
Entry Version	ALZHEIMER DIS
History Note	1998(1963); for DEMENTIA, PRESENILE and DEMENTIA, PRIMARY DEGENERATIVE SENILE use DEMENTIA 1981-1999; for DEMENTIA SENILE use DEMENTIA 1966-1999
Date of Entry	19990101
Unique ID	D000544

Tabelle 8.1: Beispieleintrag MeSH 2011 Alzheimer Krankheit

8.2 Webblog zum Thema Fettleber

Von einer Fettleber wird gesprochen, wenn mehr als 50% der Leberzellen verfettet sind. Hauptursache ist neben Überernährung vor allem chronischer Alkoholmissbrauch. Zu einer Fettleber kommt es, wenn die Leber nicht mehr in der Lage ist, ihre vielfältigen Stoffwechselleistungen auszuführen: Sie ist zuständig für die Bildung der Galle, den Fett-, Kohlenhydrat- und Eiweißstoffwechsel, die Entgiftung des Körpers, die Speicherung von Eisen und Vitaminen und den Abbau von Blutkörperchen. Zudem ist sie in der Lage, den Überschuss an bestimmten Stoffen in eine Speicherform zu überführen.

Wird dem Körper übermäßig viel Alkohol zugeführt, beginnt die Leber, diesen für den Körper giftigen Stoff abzubauen. Der Fettstoffwechsel wird zunächst vernachlässigt. Die Leber speichert die Nahrungsfette, um nach dem Alkoholabbau auf sie zurückzugreifen. Wird dem Körper kontinuierlich Alkohol zugeführt, kann die Leber ihre Aufgaben im Fettstoffwechsel nicht mehr wahrnehmen. Es wird immer mehr Fett in den Leberzellen abgelagert. Dieser Prozess der Leberzellverfettung führt zur Fettleber.

Zunächst haben Patienten keine Beschwerden. Meist wird die Fettleber erst bemerkt, wenn sich eine Fettleberhepatitis, eine Entzündung der Leber, entwickelt hat. Die Leberzellen beginnen abzusterben, die Verdauungsfunktionen des Organs sind stark eingeschränkt. Betroffene leiden unter Leistungsminderung, Übelkeit und Erbrechen. Bei weiterem Alkoholmissbrauch entwickelt sich aus der Fettleberhepatitis die Leberzirrhose. In diesem Stadium ist die Leber nicht mehr reparabel. Das Symptombild ist vielfältig: Typische Beschwerden sind zunächst ein Druckgefühl im Oberbauch, Gewichtsverlust, sowie ein ausgeprägter Ikterus: Er entsteht, wenn Abbauprodukte des Blutstoffwechsels in das Gewebe übertreten und führt zu einer Gelbfärbung der Haut und der Schleimhäute. Häufig ist er auch mit starkem Juckreiz verbunden. Neben Hautauffälligkeiten kommt es im Verlauf zu hormonellen Störungen, Pfortaderhochdruck und Leberbedingten Erkrankungen des Gehirns, die zum Tode führen können.

Die einzige oft lebensrettende therapeutische Maßnahme ist absolute Alkoholkarenz. Bei medikamentöser Behandlung ist die Vorbelastung der Leber zu berücksichtigen. Hat sich bereits eine Leberzirrhose gebildet, gilt die Prognose für den Patienten als sehr schlecht.

8.3 PubMed Artikel zum Thema Fettleber

```
<?xml version="1.0"?>
<!DOCTYPE PubmedArticleSet PUBLIC "-//NLM//DTD PubMedArticle, 1st January
    2012//EN" "http://www.ncbi.nlm.nih.gov/corehtml/query/DTD/
    pubmed_120101.dtd">
<PubmedArticleSet>
<PubmedArticle>
    <MedlineCitation Owner="NLM" Status="In-Data-Review">
        <PMID Version="1">22384276</PMID>
        <DateCreated>
            <Year>2012</Year>
            <Month>03</Month>
            <Day>02</Day>
        </DateCreated>
        <Article PubModel="Print-Electronic">
            <Journal>
                <ISSN IssnType="Electronic">1932-6203</ISSN>
                <JournalIssue CitedMedium="Internet">
                    <Volume>7</Volume>
                    <Issue>2</Issue>
                    <PubDate>
                        <Year>2012</Year>
                    </PubDate>
                </JournalIssue>
                <Title>PloS one</Title>
                <ISOAbbreviation>PLoS ONE</ISOAbbreviation>
            </Journal>
            <ArticleTitle>Preventing Phosphorylation of Sterol Regulatory
                Element-Binding Protein 1a by MAP-Kinases Protects Mice
                from Fatty Liver and Visceral Obesity.</ArticleTitle>
            <Pagination>
                <MedlinePgn>e32609</MedlinePgn>
            </Pagination>
            <Abstract>
                <AbstractText>The transcription factor sterol regulatory
                    element binding protein (SREBP)-1a plays a pivotal
                    role in lipid metabolism. Using the SREBP-1a
                    expressing human hepatoma cell line HepG2 we have
                    shown previously that human SREBP-1a is phosphorylated
                     at serine 117 by ERK-mitogen-activated protein
                    kinases (MAPK). Using a combination of cell biology
                    and protein chemistry approach we show that SREBP-1a
                    is also target of other MAPK-families, i.e. c-JUN N-
                    terminal protein kinases (JNK) or p38 stress activated
                     MAP kinases. Serine 117 is also the major
                    phosphorylation site in SREBP-1a for JNK. In contrast
                    to that the major phosphorylation sites of p38 MAPK
                    family are serine 63 and threonine 426. Functional
                    analyses reveal that phosphorylation of SREBP-1a does
                    not alter protein/DNA interaction. The identified
                    phosphorylation sites are specific for both kinase
                    families also in cellular context. To provide direct
                    evidence that phosphorylation of SREBP-1a is a
                    regulatory principle of biological and clinical
                    relevance, we generated transgenic mice expressing
```

```
            mature transcriptionally active N-terminal domain of
            human SREBP-1a variant lacking all identified
            phosphorylaton sites designed as alb-SREBP-1aP and
            wild type SREBP-1a designed as alb-SREBP-1a liver
            specific under control of the albumin promoter and a
            liver specific enhancer. In contrast to alb-SREBP-1a
            mice the phosphorylation-deficient mice develop no
            enlarged fatty livers under normocaloric conditions.
            Phenotypical examination reveales a massive
            accumulation of adipose tissue in alb-SREBP-1a but not
             in the phosphorylation deficient alb-SREBP-1aP mice.
            Moreover, preventing phosphorylation of SREBP-1a
            protects mice also from dyslipidemia. In conclusion,
            phosphorylation of SREBP-1a by ERK, JNK and p38 MAPK-
            families resembles a biological principle and plays a
            significant role, in vivo.</AbstractText>
    </Abstract>
    <Affiliation>Institute of Clinical Biochemistry and
        Pathobiochemistry, German Diabetes Center at the Heinrich-
        Heine-University Duesseldorf, Leibniz Center for Diabetes
        Research, Duesseldorf, Germany.</Affiliation>
    <AuthorList CompleteYN="Y">
        <Author ValidYN="Y">
            <LastName>Kotzka</LastName>
            <ForeName>Jorg</ForeName>
            <Initials>J</Initials>
        </Author>
        <Author ValidYN="Y">
            <LastName>Knebel</LastName>
            <ForeName>Birgit</ForeName>
            <Initials>B</Initials>
        </Author>
        <Author ValidYN="Y">
            <LastName>Haas</LastName>
            <ForeName>Jutta</ForeName>
            <Initials>J</Initials>
        </Author>
        <Author ValidYN="Y">
            <LastName>Kremer</LastName>
            <ForeName>Lorena</ForeName>
            <Initials>L</Initials>
        </Author>
        <Author ValidYN="Y">
            <LastName>Jacob</LastName>
            <ForeName>Sylvia</ForeName>
            <Initials>S</Initials>
        </Author>
        <Author ValidYN="Y">
            <LastName>Hartwig</LastName>
            <ForeName>Sonja</ForeName>
            <Initials>S</Initials>
        </Author>
        <Author ValidYN="Y">
            <LastName>Nitzgen</LastName>
            <ForeName>Ulrike</ForeName>
            <Initials>U</Initials>
```

```
            </Author>
            <Author ValidYN="Y">
                <LastName>Muller-Wieland</LastName>
                <ForeName>Dirk</ForeName>
                <Initials>D</Initials>
            </Author>
        </AuthorList>
        <Language>eng</Language>
        <PublicationTypeList>
            <PublicationType>Journal Article</PublicationType>
        </PublicationTypeList>
        <ArticleDate DateType="Electronic">
            <Year>2012</Year>
            <Month>02</Month>
            <Day>27</Day>
        </ArticleDate>
    </Article>
    <MedlineJournalInfo>
        <Country>United States</Country>
        <MedlineTA>PLoS One</MedlineTA>
        <NlmUniqueID>101285081</NlmUniqueID>
        <ISSNLinking>1932-6203</ISSNLinking>
    </MedlineJournalInfo>
    <CitationSubset>IM</CitationSubset>
    <OtherID Source="NLM">PMC3287979</OtherID>
</MedlineCitation>
<PubmedData>
    <History>
        <PubMedPubDate PubStatus="received">
            <Year>2011</Year>
            <Month>8</Month>
            <Day>11</Day>
        </PubMedPubDate>
        <PubMedPubDate PubStatus="accepted">
            <Year>2012</Year>
            <Month>1</Month>
            <Day>30</Day>
        </PubMedPubDate>
        <PubMedPubDate PubStatus="epublish">
            <Year>2012</Year>
            <Month>2</Month>
            <Day>27</Day>
        </PubMedPubDate>
        <PubMedPubDate PubStatus="entrez">
            <Year>2012</Year>
            <Month>3</Month>
            <Day>3</Day>
            <Hour>6</Hour>
            <Minute>0</Minute>
        </PubMedPubDate>
        <PubMedPubDate PubStatus="pubmed">
            <Year>2012</Year>
            <Month>3</Month>
            <Day>3</Day>
            <Hour>6</Hour>
            <Minute>0</Minute>
```

```
            </PubMedPubDate>
            <PubMedPubDate PubStatus="medline">
                <Year>2012</Year>
                <Month>3</Month>
                <Day>3</Day>
                <Hour>6</Hour>
                <Minute>0</Minute>
            </PubMedPubDate>
        </History>
        <PublicationStatus>ppublish</PublicationStatus>
        <ArticleIdList>
            <ArticleId IdType="doi">10.1371/journal.pone.0032609</
                ArticleId>
            <ArticleId IdType="pii">PONE-D-11-15758</ArticleId>
            <ArticleId IdType="pubmed">22384276</ArticleId>
            <ArticleId IdType="pmc">PMC3287979</ArticleId>
        </ArticleIdList>
    </PubmedData>
</PubmedArticle>

</PubmedArticleSet>
```

Listing 8.1: PubMed Artikel zum MeSH Term "Fettleber"

8.4 Ergebnisse des Tokenizing zu Tokenizer 3

Token	CUI	Bezeichnung
Kv1	C0290730	Kv1.3-Kaliumkanal
Kv1	C0294031	Kv1.1-Kaliumkanal
Kv1	C0295345	Kv1.5-Kaliumkanal
Kv1	C0297614	Kv1.2-Kaliumkanal
Kv1	C0299095	Kv1.4-Kaliumkanal
Kv1	C0528161	Kv1.6-Kaliumkanal
Kv1	C0663105	Kv1.2-Kaliumkanal beta-Untereinheit
Kv1	C0906484	Kv1.1-Kaliumkanal alpha-Untereinheit
Kv1	C0906576	Kv1.2-Kaliumkanal alpha-Untereinheit
Kv1	C0969358	Kv1.2'-Kanal
Kv1	C1563663	Kv1.1-Kaliumkanal beta-Untereinheit
Kanal	C0007950	Kanalinseln
Kanal	C0917707	Kanalisierung
Kanal	C0021445	Leistenkanal
Kanal	C0037922	Spinalkanal
Kanal	C0086881	Pulpakanal
Kanal	C0227411	Analkanal
Kanal	C0289061	Kv2.1-Kaliumkanal
Kanal	C0290730	Kv1.3-Kaliumkanal
Kanal	C0294031	Kv1.1-Kaliumkanal
Kanal	C0295345	Kv1.5-Kaliumkanal
Kanal	C0297614	Kv1.2-Kaliumkanal
Kanal	C0299095	Kv1.4-Kaliumkanal
Kanal	C0384156	Epithelialer Natriumkanal
Kanal	C0388155	Kv3.1-Kaliumkanal
Kanal	C0389317	TRP-Kanal
Kanal	C0528161	Kv1.6-Kaliumkanal
Kanal	C0531002	KCNQ1-Kaliumkanal
Kanal	C0662950	Kv2.2-Kaliumkanal
Kanal	C0669669	KCNQ2-Kaliumkanal
Kanal	C0669672	KCNQ3-Kaliumkanal
Kanal	C0906849	Kv3.2-Kaliumkanal
Kanal	C0910872	Zyklisch-Nukleotid-gesteuerter Kationenkanal
Kanal	C0969358	Kv1.2'-Kanal
Kanal	C1522467	Wirbelkanal
Kanal	C1563377	Kv4.2-Kaliumkanal
Kanal	C1571511	Kv4.1-Kaliumkanal
Kanal	C1571564	Kv4.3L-Kaliumkanal
Kanal	C1571639	Kv4.3-Kaliumkanal
Kanal	C1720971	Seitenlinienkanal

Tabelle 8.2: Suche nach den Token Kv1 und Kanal im UMLS Metathesaurus

Eidesstattliche Erklärung

Ich versichere, die von mir vorgelegte Arbeit selbständig verfasst zu haben.

Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten anderer entnommen sind, habe ich als entnommen kenntlich gemacht. Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben.

Die Arbeit hat mit gleichem Inhalt bzw. in wesentlichen Teilen noch keiner anderen Prüfungsbehörde vorgelegen.

Wuppertal, 12. April 2012

Zeljko Carevic