Fernuniversität Hagen

Modul 63458 - Fachpraktikum

Natural Language Processing, Information Extraction und Retrieval

WiSe 2022/23

Exposé

Automatische Erstellung einer Wissensrepräsentation aus einem medizinischen Text

eingereicht von

Anne Koch, Clara Jansen, Dietrich Tönnies

Betreuer: Prof. Dr. Hemmje
Dr. Christian Nawroth
Stephanie Heidepriem, M.Sc.

Hagen, 6. November 2022

INHALTSVERZEICHNIS

1

EINLEITUNG

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

2 MOTIVATION

Die spezielle Herausforderung unserer Praktikumsaufgabe liegt darin, automatisiert Zusammenhänge zwischen Entitäten medizinischer Texte zu finden und diese Zusammenhänge zu repräsentieren. Dem zugrunde liegt das Ziel des Dissertationsprojektes von Stefanie Heidepriem, zu erforschen, wie ein textbasierter Chatbot zur Unterstützung psychisch kranker Menschen entwickelt werden kann. Es existieren bereits Frameworks, die in der Lage sind Entitäten, also wichtige Schlüsselbegriffe, automatisiert aus Texten zu extrahieren. Dies nennt man Named Entity Recognition.

Named Entity Recognition ... [TODO, einige Worte zu NER, aktuelle Entwicklungen, Erkenntnisse im medizinischen Bereich]

Das übergeordnete Dissertationsprojekt von Stefanie Heidepriem ist unter anderem an das Projekt MENHIR angelehnt. Dieses Projekt dient der Erforschung von Konversationstechnologien, die psychisch kranke Menschen ünterstützen sollen. In diesem Zusammenhang wird auch ein MENHIR-Chatbot entwickelt, der psychisch kranken Menschen personalisierte Unterstützung und hilfreiche Bewältigungsstrategien bieten soll.

3 PROBLEMBESCHREIBUNG

Es soll eine Konsolenapplikationen entwickelt werden, die medizinische Texte analysiert und das darin enthalten Wissen strukturiert in einer XML-Datei dokumentiert oder mittels eines RDF-Modells visualisiert werden. Die Konsolenapplikation soll in Python programmiert werden, wobei für die Textanalyse Klassen und Methoden der Open-Source-Bibliothek spaCy genutzt werden sollen.

Die medizinischen Texte enthalten z.B. Aussagen zu Krankheiten (z.B. Depression) und zählen deren Symptome (z.B. Motivationsverlust) auf. Aufgabe der Konsolenapplikation ist es dann, Schlüsselbegriffe im Text zu finden und miteinander in Bezug zu setzen, indem z.B. Symptome den ihnen zugrunde liegenden Krankheiten zugeordnet werden. Die von spaCy zur Verfügung gestellte Funktionalität besteht aus einer Pipeline von Analyse-Komponenten, die nacheinander auf den zu analysierenden Text angewendet werden. Diese Komponenten dienen dazu, Texte in einzelne Sätze zu zerlegen und die Sätze grammatikalisch zu analysieren.

Neben der grammatikalischen Analyse besteht eine wesentliche Aufgabe von spaCy darin, Schlüsselbegriffe zu finden und zu kategorisieren. Diese Art der Analyse wird allgemein als Named Entity Recognition (NER) bezeichnet. Bei spaCy übernimmt diese Aufgabe der regelbasierte EntityRuler oder der auf statistischen Modellen basierende EntityRecognizer. Eine weitere Komponente ist der EntityLinker, mit dem Begriffe eindeutig den in einer Wissensbasis gespeicherten Entitäten zugeordnet werden können. Auch der EntityLinker basiert auf statistischen Modellen und wird mit Beispielsätzen trainiert.

Die Standard-NER Funktionalität von spaCy erkennt medizinische Begriffe nur unzureichend. Es ist daher notwendig, eine Datenbank mit medizinischen Begriffen und Kategorien zusammenzustellen, die für das Training der NER-Komponenten von spaCy verwendet werden kann.

Die auf dem Campus des National Institute of Health (NIH) im US-Bundestaat Maryland angesiedelte National Library of Medicine stellt mit dem Unified Medical Language System (UMLS) ein mächtiges Werkzeug für die Textanalyse zur Verfügung. Teil von UMLS ist der sogenannte Metathesaurus, der aus einer Vielzahl von Thesauri unterschiedlicher Organisationen zusammengestellt wird. Zu diesen Thesauri gehört der vom National Library of Medicine entwickelte Medical Subject Heading (MeSH)-Thesaurus. MetaMap und das weniger umfangreiche MetaMapLite sind eigene Entity-Recognition-Werzeuge der National Library of Medicine. Die zugrundeliegende Datenbasis dieser Werkzeuge sollen im Rahmen dieses Praktikums dafür verwendet werden, die Named Entity Recognition-Komponenten von spaCy zu trainieren.

Basierend auf den gefundenen Entitäten soll das Python-Programm in der Lage sein, Begriffe wie Krankheiten und Symptome richtig zuzuordnen. Hierzu müssen Aussagesätze, Fragen, Aufzählungen und Überschriften richtig gedeutet werden.

- Medizinische Begriffe, die mithilfe von Named Entity Recognition gefunden werden, müssen zueinander in Zusammenhang gebracht werden.
- Es gibt keine vorgegebene / einheitliche Struktur für die Wissensrepräsentation von Entity Linking.

4 FORSCHUNGSFRAGEN

Dieser Abschnitt beschreibt die Forschungsfragen, die sich aus dem in Abschnitt 3 angeführten Problemen ableiten.

FF 1 Wie kann ein medizinischer Text vorverarbeitet werden, so dass relevante Aussagen (Überschriften, Aufzählungen, Leerzeilen und Dialoge) automatisch erkannt werden?

FF 2 Wie kann ein medizinisches Fachvokabular über depressive Erkrankungen automatisiert in eine maschinenlesbare Form überführt werden?

FF 3 Wie kann eine Wissensrepräsentation über einen gegebenen Text maschinenlesbar erstellt werden?

5 FORSCHUNGSMETHODE

In dieser Arbeit kommt die in der Forschung zu Informationssystemen etablierte Methode von Nunamaker, Chen und Purdin [NCP90] zum Einsatz, die einen strukturierten Rahmen zur Durchführung von Forschungsarbeiten bereitstellt.

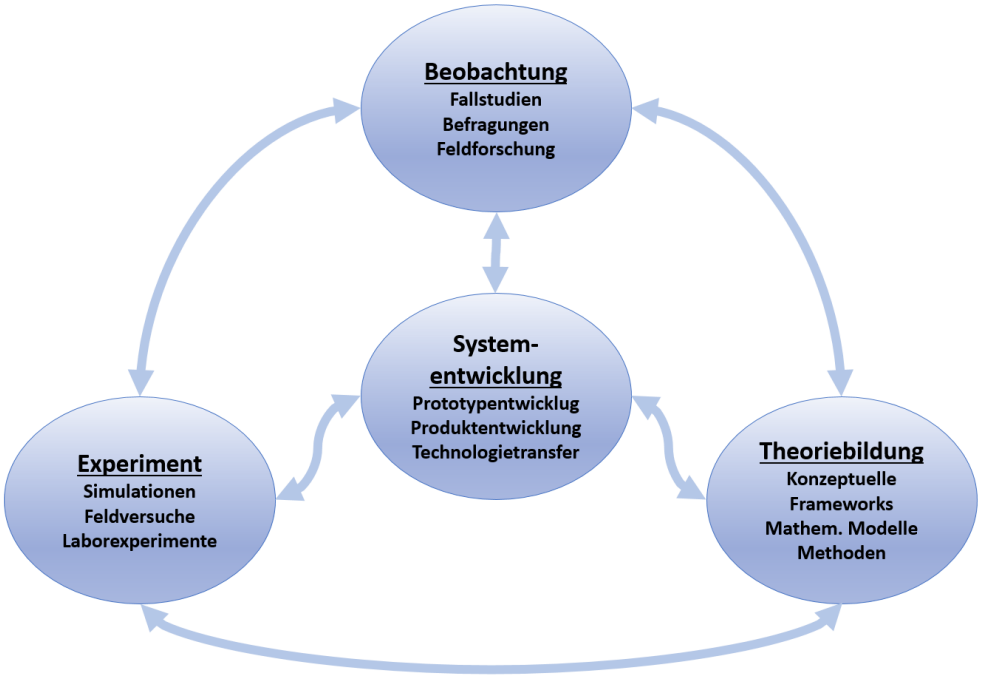

Abbildung 5.1: Forschungsmethode nach [NCP90]

Die Forschungsmethode umfasst die folgenden in Abbildung 5.1 dargestellten Phasen:

BEOBACHTUNG: Insbesondere wenn der Untersuchungsgegenstand relativ unbekannt ist, werden Fallstudien, Feldversuche oder Umfragen durchgeführt, um ein "Gefühl"für den Aufwand zu erhalten. Auf dieser Grundlage können dann konkrete Hypothesen erstellt werden, die durch Experimente geprüft werden. Zusätzlich dazu wird in dieser Phase die Literaturrecherche und Ermittlung des Standes von Wissenschaft und Technik durchgeführt.

THEORIEBILDUNG: In dieser Phase werden neue Ideen, Konzepte Methoden oder Modelle entwickelt. Diese Theorien beschreiben das Systemverhalten allgemein, haben jedoch wenig praktische Bedeutung für die Zieldomäne. Sie können aber genutzt werden, um Forschungshypothesen aufzustellen, die Planung von Experimenten zu unterstützen und systematische Beobachtungen durchzuführen.

SYSTEMENTWICKLUNG: Diese Phase besteht aus fünf Teilen:

- Konzeptentwurf
- Erstellung der Systemarchitektur
- Erstellung von Prototypen
- Produktentwicklung
- Technologietransfer

EXPERIMENT: In dieser Phase werden die gefundenen Theorien und entwickelten Systeme evaluiert. Die Ergebnisse der Experimente können genutzt werden, um die Theorien zu weiterzuentwickeln und die Systeme zu verbessern.

Obwohl die Phasen in der Methode keine vorgegebene Reihenfolge haben, sondern sich gegenseitig beeinflussen, wird in der vorliegenden Arbeit die oben angeführte Abfolge verwendet.

6 FORSCHUNGSZIELE

Dieser Abschnitt beschreibt die Forschungsziele, die sich aus den in Abschnitt 4 angeführten Fragen ableiten.

FZ 1 Wie kann ein medizinischer Text vorverarbeitet werden, so dass relevante Aussagen (Überschriften, Aufzählungen, Leerzeilen und Dialoge) automatisch erkannt werden?

- FZ 1.1 Aufbau und Struktur medizinischer Texte (Observation)
- FZ 1.2 Theoriebildung zur Vorverarbeitung medizinischer Texte (Theoriebildung)
- FZ 1.3 Implementierung zur Vorverarbeitung medizinischer Texte (Implementierung)
- FZ 1.4 Evaluation zur Vorverarbeitung medizinischer Texte (Evaluation)

FZ 2 Wie kann ein medizinisches Fachvokabular über depressive Erkrankungen automatisiert in eine maschinenlesbare Form überführt werden?

- FZ 2.1 Überblick über depressive Erkrankungen (Observation)
- FZ 2.2 Automatisierung durch NLP (Observation)
- FZ 2.3 Theoriebildung zur Überführung medizinischen Fachvokabulars in maschinenlesbare Form zur weiteren Verarbeitung durch NLP (Theoriebildung)
- FZ 2.4 Implementierung zur Überführung medizinischen Fachvokabulars in maschinenlesbare Form zur weiteren Verarbeitung durch NLP (Implementierung)
- FZ 2.5 Evaluation zur Überführung medizinischen Fachvokabulars in maschinenlesbare Form zur weiteren Verarbeitung durch NLP (Evaluation)

FZ 3 Wie kann eine Wissensrepräsentation über einen gegebenen Text maschinenlesbar erstellt werden?

- FZ 3.1 Wissensrepräsentation mittels RDF, RDFS (Observation)
- FZ 3.2 Theoriebildung zur Wissensrepräsentation (Theoriebildung)
- FZ 3.3 Implementierung zur Wissensrepräsentation (Implementierung)
- FZ 3.4 Evaluation zur Wissensrepräsentation (Evaluation)

7 LÖSUNGSANSATZ

Arbeitspakete:

- Ein medizinisches Vokabular muss eingebunden werden um Named Entities identifizieren zu können (MEsH).
- Eine Named Entity Recognition muss für den zu analysierenden Text ausgeführt werden.
- (Eine Knowledge Base für das Entity Linking muss erstellt werden.)
- Eine manuelle Wissensrepräsentation basierend auf dem zu analysierenden Text muss erstellt werden.
- Die automatisch erstellte Wissensrepräsentation muss mit der manuell erstellten Wissensrepräsentation verglichen werden.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

8

STAND DER WISSENSCHAFT

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

9 ARBEITS- UND ZEITPLAN

		Beginn	ETA
1	Gesamtprojekt	15.11.22	28.03.23
2	PE3 Prototypische Softwareentwicklung	15.11.22	16.12.22
3	Meilenstein 1	15.11.22	16.12.22
4	Modellierung und Theoriebildung	15.11.22	16.12.22
5	Einleitung	15.11.22	
6	Stand der Technik		16.12.22
7	FF1, FF2, FF3	15.11.22	16.12.22
8	Meilenstein 2	17.12.22	17.02.23
9	Systementwicklung	17.12.22	17.02.23
10	Implementation	17.12.22	17.02.23
11	FF1, FF2, FF3	17.12.22	17.02.23
12	Meilenstein 3	18.02.23	20.03.23
13	Experiment	18.02.23	20.03.23
14	Evaluation	18.02.23	20.03.23
15	FF1, FF2, FF3	18.02.23	20.03.23
16	PE4, Präsentation der Ergebnisse	21.03.23	28.03.23
17	Präsentation	21.03.23	25.03.23
18	Kolloquium	25.03.23	28.03.23

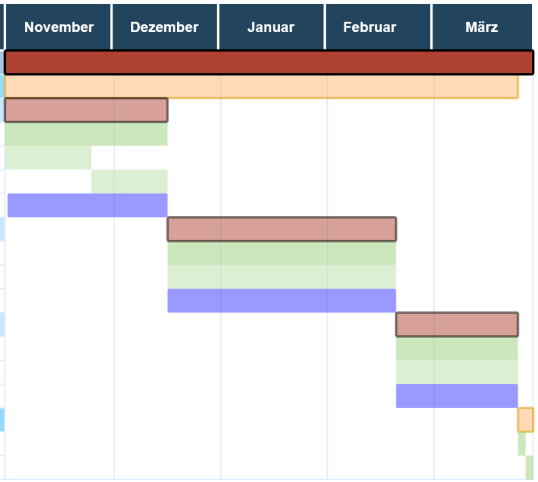

Abbildung 9.1: Arbeits- und Zeitplan

[NCP90] Jay F. Nunamaker, Minder Chen und Titus D.M. Purdin. “Systems Development in Information Systems Research”. In: *Journal of Management Information Systems* 7.3 (Dez. 1990). Publisher: Routledge _eprint: https://doi.org/10.1080/07421222.1990.11517898, S. 89–106. ISSN: 0742-1222. DOI: 10.1080/07421222.1990.11517898. URL: https://doi.org/10.1080/07421222.1990.11517898 (besucht am 02. 11. 2022).