Universität Leipzig
Medizinische Fakultät
Institut für Medizinische Informatik, Statistik und Epidemiologie

QUESTION ANSWERING AUF SNIK MIT EINEM LANGUAGE MODEL

MASTERARBEIT

zur Erlangung des akademischen Grades

Master of Science
(M. Sc.)

vorgelegt von

Paul Keller
Studiengang Medizininformatik M. Sc.

Leipzig, den 31.09.2023

[11. März 2023 at 13:36 – classicthesis v4.6]

AUTOR:
Paul Keller
Geboren am 23.05.1998 in Leipzig, Deutschland

TITEL:
Question Answering auf SNIK mit einem Language Model

INSTITUT:
Institut für Medizinische Informatik, Statistik und Epidemiologie
Medizinische Fakultät Universität Leipzig

REFERENT:
Prof. Dr. Alfred Winter

BETREUER:
Konrad Höffner

ABSTRAKT

DANKSAGUNG

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

AKRONYME

1 EINLEITUNG

1.1 GEGENSTAND

Eine effektive und effiziente Informationsgewinnung bildet einen fundamentalen Bestandteil einer qualitativ hochwertigen klinischen Praxis in der Medizin. Bei jeder medizinischen Handlung werden große Mengen an Informationen genutzt und generiert, sei es zur Grundlage einer Diagnose oder zur Dokumentation des Behandlungsprozesses. Die strukturierte und klassifizierte Speicherung sowie Wiedergabe dieser Informationen stellt einen fortwährenden Entwicklungsprozess dar und ist Gegenstand aktueller Forschung.

Die Digitalisierung der Medizin ist ein ausgedehntes Themenfeld mit stetig wachsender Notwendigkeit. Die Medizinische Informatik beschreibt passend dazu „die Wissenschaft der systematischen Erschließung, Verwaltung, Aufbewahrung, Verarbeitung und Bereitstellung von Daten, Informationen und Wissen in der Medizin und im Gesundheitswesen. [...]"(gmds, 2023). In diesem Kontext wird die Bedeutung der Entwicklung und Implementierung effektiver Informationssysteme und Technologien zur Unterstützung der klinischen Praxis immer größer.

In der Lehre wird die Praxis der medizinischen Informatik durch umfassende Literatur wie zum Beispiel in Winter u. a. (2011) unterstützt. Zur Strukturierung dieser Literatur existiert die Ontologie SNIK (Jahn u. a., 2019), ein semantisches Netzwerk klassifiert in das SNIK Metamodell und Teil des Instituts für Medizinische Informatik, Statistik und Epidemiologie (IMISE, 2023) an der Universität Leipzig. Durch die Verwendung dieses Netzwerkes wird eine systematische Darstellung von Rollen, Entitäten und Funktionen in der Medizinischen Informatik ermöglicht, losgelöst von den Definition der zugrunde liegenden Literaturquellen.

SNIK kann über diverse Zugriffsmöglichkeiten über das Internet durchsucht werden. Der RDF-Browser (SNIK, 2023b) bietet eine rudimentäre Möglichkeit, auf die SNIK-Ontologie zuzugreifen, während der SPARQL-Endpunkt (SNIK, 2023c) einen Query-Editor bereitstellt, der es ermöglicht, graphenbasierte Anfragen an das Resource Description Framework (RDF) von SNIK zu formulieren. Ferner erlaubt eine Graphvisualisierung (SNIK, 2023a) eine Darstellung der Informationsendpunkte und ihrer Vernetzung als Graphen. Schließlich stellt

QAnswer (QAnswer, 2023) eine Möglichkeit dar, Fragen in englischer Sprache über das Netzwerk zu stellen und in tabellarischer Form Antworten zu erhalten.

1.2 PROBLEMSTELLUNG

Zur Unterstützung der Lehre ist die Nutzung der in 1.1 genannten Zugriffsarten subomptimal. Sowohl für Lehrende als auch Studierende bieten diese Arten keine intuitive Benutzung oder Ausgabe der Informationen an.

Der RDF-Browser bietet, limitiert durch seine rudimentäre Art, sowohl eine unintuitive Suche als auch Darstellung der Informationen in SNIK. Die Visualisierung des Graphen bietet zwar eine annehmliche Darstellung, wird jedoch bei komplexeren Fragen schwierig in ihrer Navigation und unübersichtlich in ihrer Darstellung. Der SPARQL Editor als programatischer Endpunkt enthält sowohl die Hürde der SPAR Query Language als Voraussetzung seiner Nutzung, als auch keine übersichtliche Darstellung der Ergebnisse. Question Answerung mit Hilfe von QAnswer glänzt durch ihren Ansatz der Nutzung von natürlicher Sprache als Fragestellung. Jedoch ist auch hier die Darstellung der Ergebnisse, wenn sie denn korrekt ist, nicht in natürlicher Form.

Jede Zugriffsart hat ihre Existensberechtigung und unterschiedlichste Anwendungsfälle. Doch eine Interaktion mit SNIK und den zugrunde liegenden Literaturquellen von Studierenden und Lehrenden ist stets limitiert in ihrer natürlichen Nutzung oder Darstellung der Daten.

Durch diese fehlenden Zugriff ist das SNIK Projekt unüblich als Informationsbasis in der Lehre, obwohl dessen Intention jene Lehre erheblich verbessern würde. Häufige semantische Probleme wie Synonyme und Homonyme für Begriffe als auch eine Zusammenfassung der in den Literaturquellen vorhandenen Informationen sind durch SNIK gegeben, können aber nicht von Studierenden einfach genutzt werden.

- Problem P1: Aktuelle Zugriffsarten auf SNIK und dessen Literaturquellen bieten keinen einfachen und intuitiven Einstieg
- Problem P2: Wiedergabe von Informationen aus SNIK und dessen Literaturquellen ist unübersichtlich und erfordert zusätzliches Hintergrundwissen zum Verständnis

1.3 MOTIVATION

Allgemeines Wissen, Informationen und Daten werden in heutiger Zeit über Google und Wikipedia gefunden. Die Suche nach Fragen durch Google und die korrekte Beantwortung jener durch einen kurzen Ausschnitt aus der Definition des Begriffs in Wikipedia bieten dem normalen Nutzer eine einfache und schnelle Möglichkeit, seine/ihre Fragen in natürlicher Sprache zu formulieren und Antworten in natürlicher Sprache zu erhalten.

Dies erfordert jedoch enorme Rechenleistungen und große Datensätze um konkretes und korrektes Wissen wiederzugeben. Weitergehend sind Wissen und Informationen über die Medizinische Informatik selten im selben Detailgrad in diesen Ressourcen vorhanden, wie sie in den Literaturquellen von SNIK vorliegen. Zur Unterstützung von Lehrenden und Studierenden soll diese Arbeit eine Möglichkeit bieten, ähnlich dem Prozess der gerade beschriebenen normalen Suche im Internet, Wissen und Informationen aus den Literaturquellen zu extrahieren, welche auch als Basis für die Erstellung des SNIK-Projektes genutzt wurden.

Dies ermöglicht jenen Nutzenden die Problematiken bei dem Umgang mit SNIK zu umgehen und bietet eine weitere Methodik um digitale Bildung besser in den Arbeitsfluss einer Universität zu integrieren.

1.4 ZIELSETZUNG

Den in 1.2 gezeigten Problemen P_i werden Ziele Z_i dieser Arbeit zugeordnet.

- Ziel Z1: Zugriff auf Literaturquellen von SNIK mit Hilfe eines ChatBots
- Ziel Z2: Wiedergabe von Wissen aus den Literaturquellen von SNIK in natürlicher Sprache beispielhaft gezeigt an Hand von Winter u. a. (2011) als Wissensressource

1.5 AUFGABENSTELLUNG

Die in 1.4 genannten Ziele Z_i werden durch die hier aufgeführten Aufgaben A_i gelöst.

- Aufgabe zu Ziel Z1
  - Aufgabe A1.1: Implementierung einer Webseite, welche natürlich gestellte Fragen entgegen nimmt

  - Aufgabe A1.2: Nutzung eines General Pre-trained Transformers (GPT) zur Bewertung, Verständis und Beantwortung der gegeben Fragen

- Aufgabe zu Ziel Z2
  - Aufgabe A2.1: Erstellung eines Fragenkatalog aus der Literaturquelle
  - Aufgabe A2.2: Fein-Tuning des GPT-Modells auf die Literaturquelle
  - Aufgabe A2.3: Analyse und Bewertung der ausgegebenen Daten

1.6 AUFBAU DER ARBEIT

Kapitel 1 beschreibt das grundlegende Umfeld dieser Arbeit, formuliert exitierende Probleme und Anforderungen, bietet Ziele zur Lösung dieser Probleme an und gibt Aufgaben, zur Umsetzung dieser Ziele. In Kapitel 2 werden Grundlagen gelegt zum Verständnis der in dieser Arbeit verwendeten Technologie, während in Kapitel 3 der aktuelle Stand der Forschung zusammengefasst wird. Kapitel 4 umfasst Lösungsstrategien der in 1.2 formulierten Probleme mit einer anschließenden Beschreibung der Umsetzung dieser Lösungen in Kapitel 5. Die Ergebnisse dieser Arbeit werden in Kapitel 6 präsentiert und in Kapitel 7 zusammengefasst diskutiert. Zusätzlich gibt Kapitel 7 einen Ausblick dieser Arbeit.

2

GRUNDLAGEN

3

STAND DER FORSCHUNG

4

LÖSUNGSANSATZ

5 AUSFÜHRUNG DER LÖSUNG

6 ERGEBNISSE

7 DISKUSSION

ZUSAMMENFASSUNG

LITERATUR

IMISE (2023). *Institut für Medizinische Informatik, Statistik und Epidemiologie*. URL: https://www.imise.uni-leipzig.de/Institut (besucht am 09. 03. 2023).

Jahn, Franziska, Konrad Höffner, Birgit Schneider, Anna Lörke, Thomas Pause, Elske Ammenwerth und Alfred Winter (2019). „The SNIK Graph: Visualization of a Medical Informatics Ontology". In: *MedInfo 2019, The 17th World Congress of Medical and Health Informatics, Lyon*.

QAnswer (2023). *Question Answerung*. URL: https://app.qanswer.ai/public-share?kb=SNIK_BB&type=graph&user=kirdie (besucht am 09. 03. 2023).

SNIK (2023a). *Graphvisualisierung*. URL: https://www.snik.eu/graph/ (besucht am 09. 03. 2023).

SNIK (2023b). *RFD-Browser*. URL: https://www.snik.eu/ontology/ (besucht am 09. 03. 2023).

SNIK (2023c). *SPARQL Query Editor*. URL: https://www.snik.eu/sparql/ (besucht am 09. 03. 2023).

Winter, Alfred, Reinhold Haux, Elske Ammenwerth, Birgit Brigel, Nils Hellrung und Franziska Jahn (2011). *Health Information Systems: Architectures and Strategies*. Health Informatics. Springer London. ISBN: 9781849964418. URL: https://books.google.de/books?id=RzvmrgwCWncC.

gmds (2023). *Definition Medizinische Informatik*. URL: https://www.gmds.de/aktivitaeten/medizinische-informatik/ (besucht am 07. 03. 2023).

APPENDIX

ERKLÄRUNG

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

Leipzig, 31.09.2023

Paul Keller