Entwicklung eines Dialog-Konzeptes für einen KI-basierten Chatbot im Hochschulbereich

Maximilian Hönig (B.Sc.)

Technische Hochschule Mittelhessen

Fachbereich MND
Wilhelm-Leuschner-Str. 13
61169 Friedberg
E-Mail: maximilian.hoenig@mnd.thm.de

Prof. Dr. Harald Ritz

Technische Hochschule Mittelhessen

Fachbereich MNI
Wiesenstraße 14
35390 Gießen
E-Mail: harald.ritz@mni.thm.de

Kategorie

Abschlussarbeit

Schlüsselwörter

KI, Chatbot, Digitale Assistenten, Regeldatenbank, NLP

Zusammenfassung

Die Technische Hochschule Mittelhessen (THM) setzt einen Chatbot namens *„Winfy“* zur Beantwortung von Fragen im Kontext von Prüfungsangelegenheiten der Studiengänge B.Sc. und M.Sc. Wirtschaftsinformatik ein (vgl. Ritz/Tansel (2023) und Ludwig/Ritz (2023); URL: https://feedback.mni.thm.de/winfy/). Der Dialog zwischen diesem und dem Nutzer ist bisher statisch. Der Chatbot *„Winfy“* antwortet als FAQ-Bot nämlich direkt auf Fragen. Er stellt zum Beispiel keine Rückfragen und besitzt kein Gedächtnis (vgl. Ritz/Tansel 2023). Ziel der Masterarbeit ist es, allgemeine Möglichkeiten zur Optimierung des Chatbots zu identifizieren und umzusetzen, insbesondere aber auch die Fähigkeiten des Bots im Dialog mit dem Nutzer noch besser zu gestalten.

Eine Analyse externer Chatbots half dabei, Verbesserungspotentiale zu erkennen. Die analysierten Chatbots sollten dabei möglichst ähnlich zum Chatbot *„Winfy“* sein und stammten ebenfalls aus Hochschulen, aber auch aus Unternehmen. Die wichtigsten erkannten Potentiale waren dabei:

- Einführung/Verbesserung von Smalltalk
- Fähigkeit, Rückfragen zu stellen
- Einführung von Themen/Kategorien
- Unterstützung von Fremdsprachen
- Kürzere/prägnantere Antworten

Die meisten Punkte erscheinen selbsterklärend, aber warum sollte ein Chatbot, der nicht für das Halten von Konversationen konzipiert wurde, trotzdem in einem gewissen Maß Smalltalk beherrschen? Chatbots werden in Umfragen häufig als zu unpersönlich beschrieben. Smalltalk ermöglicht eine erste emotionale Annäherung und sollte den Bot somit persönlicher wirken lassen (vgl. Adam et al. 2021).

Für die konkrete Umsetzung ist die technische Grundlage entscheidend. Der Chatbot *„Winfy“* nutzt für die Spracherkennung Embeddings und verwendet zur Beantwortung der Fragen eine Regeldatenbank. In dieser sind Beispielfragen mit den jeweils dazugehörigen Antworten hinterlegt. Embeddings sind mathematische Repräsentationen von Wörtern oder Sätzen in Form von Vektoren. In diesem Fall werden alle Beispielfragen der Regeldatenbank in Vektoren umgewandelt. Stellt der Nutzer eine Frage, wird auch diese in einen Vektor transformiert. Anschließend wird die Beispielfrage in der Datenbank gesucht, die der Nutzerfrage am ähnlichsten ist. Im Vektorraum liegen ähnliche Begriffe oder Sätze nah beieinander. Die Ähnlichkeit zwischen zwei Sätzen, und damit zwischen ihren Vektoren, wird hier mit der Cosinus-Ähnlichkeit berechnet. Die Berechnung erfolgt, indem das Skalarprodukt der beiden Vektoren durch das Produkt ihrer Längen geteilt wird. Die hinterlegte Antwort der Beispielfrage mit der größten Ähnlichkeit wird anschließend ausgegeben.

Aufgrund der Leistungsfähigkeit von generativen Transformer-Modellen wurde die Implementierung eines solchen erwägt (vgl. Yenduri et al. 2023). Das Generieren von Antworten wurde jedoch allgemein verworfen. Der Chatbot *„Winfy“* beantwortet unter anderem Fragen über Prüfungsangelegenheiten des Studiengangs, wobei eine korrekte, in einem bestimmten Wortlaut gegebene Antwort nötig ist. Dies könnte mit generativen Modellen nicht gänzlich sichergestellt werden.

Zur Einführung von Smalltalk wurden entsprechende Regeln in der Regeldatenbank hinterlegt. Eine Bearbeitung und Aufteilung der vorhandenen Regeln ermöglicht es, feingranularere Antworten zu geben. Regeln können nun auf eine oder mehrere Kategorien verweisen, die in einer separaten Tabelle gespeichert sind. Der Nutzer kann über eine Liste eine Kategorie auswählen. In diesem Fall werden zur Beantwortung der

Frage des Nutzers nur Regeln betrachtet, die die ausgewählte Kategorie besitzen. Äquivalentes gilt für Fragenvorschläge, die der Chatbot dem Nutzer macht.

Zur Einführung von Rückfragen wurden zwei regelbasierte Vorgehensweisen implementiert, welche nachfolgend als ‚einfache' und ‚multiple' Rückfragen genannt werden. Dazu wurde den Einträgen in der Regeldatenbank ein Attribut hinzugefügt, das eben diese Rückfrage speichert. Diese muss mit ‚ja' oder ‚nein' beantwortbar sein, also z.B. „Möchten Sie wissen, wo Sie das Modulhandbuch finden?". Bei dem Vorgehen für einfache Rückfragen wird die hinterlegte Rückfrage ausgegeben, insofern die Cosinus-Ähnlichkeit nur einen unbefriedigenden Wert erreicht - oder sprichwörtlich, wenn der Chatbot sich unsicher ist. Der Nutzer kann mit ja oder nein antworten und erhält entweder die Antwort oder eine standardisierte Entschuldigung. Eine multiple Rückfrage wird hingegen gestellt, wenn mehrere Regeln eine ähnliche Cosinus-Ähnlichkeit besitzen, also mehrere Antworten in Frage kommen. In diesem Fall werden die Rückfragen in einer Nachricht gesammelt und dem Nutzer vorgestellt. Durch eine Nummerierung der Rückfragen ist der Nutzer über die Angabe der entsprechenden Zahl in der Lage sich die Frage auszuwählen, die er gerne beantwortet haben möchte. Zur Bestimmung der infrage kommenden Antworten wurde eine konfigurierbare Variable eingeführt, die die größte zulässige Abweichung von der höchsten gefundenen Cosinus-Ähnlichkeit beschreibt. Befindet sich innerhalb dieser Abweichung keine weitere Regel, wird die Antwort der Regel mit der höchsten Cosinus-Ähnlichkeit normal ausgegeben.

Die Änderungen konnten vollständig implementiert werden. Der Smalltalk und die einfachen Rückfragen ermöglichen einen natürlicheren Gesprächsverlauf, der den Bot intelligenter wirken lässt. Durch die Aufteilung vorhandener Regeln kann der Chatbot *„Winfy"* zukünftig präziser auf Fragen der Nutzer antworten. Gleichzeitig wird die Regelbasis dadurch homogener. Einzelne Regeln unterscheiden sich also weniger, und es entsteht dadurch eine Verwechselungsgefahr zwischen ihnen. Das Sprachmodell hat es somit schwerer, die Fragen korrekt zu klassifizieren. Dem entgegen wirkt die Implementierung der multiplen Rückfragen. Selbst wenn der Bot nicht in der Lage wäre, die Regeln korrekt auseinander zu halten, würde er in diesem Fall die möglichen Optionen dem Nutzer präsentieren und dieser kann die gewünschte Frage auswählen. Durch die Kategorien wird, wie bereits erwähnt, sowohl beim Beantworten von Fragen als auch beim Geben von Fragenvorschlägen die Regelbasis vorher nach der gewählten Kategorie gefiltert.

Literatur

Adam, M.; Wessel, M.; Benlian, A. (2021): AI-based chatbots in customer service and their effects on user compliance. Electron Markets, vol 31, S.427-455 URL: https://link.springer.com/article/10.1007/s12525-020-00414-7

Ludwig, S.; Ritz, H. (2023): Entwicklung einer plattformunabhängigen Chatbot-Frontend-Anwendung. Anwendungen und Konzepte in der Wirtschaftsinformatik (AKWI), Nr. 18, S.170-171, DOI: https://doi.org/10.26034/lu.akwi.2023.n18

Ritz, H.; Tansel, D. (2023): Entwicklung eines KI-basierten FAQ-Chatbots für die Hochschule im Bereich Prüfungsangelegenheiten. Anwendungen und Konzepte in der Wirtschaftsinformatik (AKWI), Nr. 17 , S.81-92 URL: https://akwi.hswlu.ch/article/view/3972

Yenduri, G.; Murugan, R.; Govardanan, C.; u.a. (2023): GPT (Generative Pre-trained Transformer) – A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. URL: https://arxiv.org/abs/2305.10435