VERGLEICHENDE EVALUATION VON ZWEI ANSÄTZEN FÜR EIN QUESTION-ANSWERING SYSTEM

Katja Schreiber, Stefan Hillman

Technische Universität Berlin
k.schreiber@tu-berlin.de

Zusammenfassung: Dieser Beitrag beschreibt einen technischen und empirischen Vergleich zwischen zwei Sprach-Interaktionssystemen. Bei einem System (SCIO-System) handelt es sich um ein regelbasiertes System und bei dem anderen System (BERT-System) um ein sprachmodellbasiertes System. Eine technische Evaluation betrachtet die quantitative korrekte Beantwortung von Fragen der Nutzenden. Dabei scheidet das BERT-System in Summe besser ab als das SCIO-System. In einer empirischen Evaluation haben die Nutzer:innen die Zugänglichkeit und Nützlichkeit der Antworten bewertet. Bei dieser Evaluation schnitt das SCIO-System besser ab. Beide Systeme zeigten in unterschiedlichen Bereichen ihre Stärken und Schwächen, weshalb für eine optimale Nutzung die einzelnen Systeme zum jetzigen Zeitpunkt nicht geeignet sind. Jedoch zeigen beide Systeme Potenzial, um mit weiteren Verbesserung und weiterer Forschung die bisherigen Ergebnisse Qualität der Antworten zu verbessern.

1 Einleitung

In dieser Arbeit sollen zwei interaktive sprachbasierte Assistenzsysteme mithilfe einer technischen und einer empirischen Evaluation bzgl. ihrer Usability verglichen werden. Beide Systeme beantworten Fragen von Nutzer:innen über Versicherungsprodukte. Bei dem einen System (SCIO-System) handelt es sich um ein regelbasiertes System, das mithilfe des Alexa Skills Kit (ASK) entwickelt wurde. Das andere System (BERT-System) basiert auf einer nachtrainierten Version des Sprachmodells Distil-BERT, das speziell auf die NLP-Aufgabe des „Question Answering“ trainiert wurde. Durch die technische Evaluation haben wir die Quantität der richtigen Antworten beider Systeme verglichen. Bei der empirischen Evaluation haben wir den Mehrwert der Antworten für Nutzer:innen verglichen.

2 Verwandte Arbeiten

Für die Implementierung des BERT-Systems wurde das bidirektionale auf Transformern basierende Sprachmodell DistilBERT, das auf dem Sprachmodell BERT basiert, benutzt. BERT steht für „Bidirectional Encoder Representation from Transformers“ und wurde 2018 von Jacob Devlin et.al. veröffentlicht [1]. DistilBERT ist im Unterschied zu BERT um 60 % schneller und erhält dabei 95 % der Performanz [2]. Bei DistilBERT handelt es sich um ein vortrainiertes Sprachmodell, das durch sogenanntes „fine-tuning“ auf verschiedene Aufgaben trainiert werden kann. Zum Beispiel wurde das Sprachmodell mit dem „Standford Question Answering Dataset“ (SQUAD) für das Beantworten von Fragen anhand eines gegebenen Kontextes nachtrainiert. Bei SQUAD 1.1 handelt es sich um eine Datenbank von mehr als 100.000 Frage-Antwort-Paaren [3]. Ein mit dieser Datenbank trainiertes Modell ist frei zugänglich über die Huggingface-Bibliothek unter dem Namen „DistilBertForQuestionAnswering“ verfügbar [2].

3 Implementierungen

Im Folgenden werden der Aufbau und die Funktionsweise des regelbasierten und sprachmodellbasierten Systems kurz beschrieben.

3.1 Regelbasiertes System

Bei dem SCIO-System handelt es sich um einen Alexa-Skill der Fragen zu Versicherungsbedingungen beantwortet. Das SCIO-System wurde von der id-Fabrik in Auftrag gegeben und von der OEV Berlin entwickelt. Nutzer:innen können dem System eine Frage stellen, die daraufhin von dem System transkribiert und beantwortet wird. Anhand von spezifischen Fragemustern bestimmt das SCIO-System eine Antwort aus einer hinterlegten Datenbank. Zusätzlich zu der Antwort sind in der Datenbank Information für die Wiedergabe an den/die Nutzer:in hinterlegt. Aus den vorhandenen Informationen generiert der Alexa-Skill eine Bildschirmoberfläche mit der Antwort und gibt den Text (und in einigen Fällen zusätzliche Informationen per Audioausgabe) über die Sprachschnittstelle wieder.

3.2 Sprachmodellbasiertes System

Der Interaktionsablauf beim BERT-System ist dem beim SCIO-System ähnlich, aber die Verarbeitung der Frage ist beim BERT-System unterschiedlich. Nutzer:innen stellen dem System eine Frage, die es transkribiert. Zusätzlich dazu wird die Frage vom Deutschen ins Englische übersetzt. Dieser Schritt ist nötig, da zum Zeitpunkt der Implementierung des BERT-Systems ein englisches Sprachmodell ausgesucht wurde, das die Zielsetzung der Implementierung am besten erfüllt. Die übersetzte Frage und die ins Englische übersetzten Versicherungsbedingungen, werden an das BERT-Modell übergeben, welches anhand der Versicherungsbedingungen mögliche Antworten auf die Frage generiert. Die Antworten gibt es mit einem Score, der jeweils die Wahrscheinlichkeit der Richtigkeit einer Antworten angibt, zurück. Die Antworten werden zurück ins Deutsche übersetzt und in absteigender Reihenfolge des Scores an die Nutzer:innen zurückgegeben.

4 Evaluationsmethodik

Zum Vergleich der beiden Systeme wurden zwei Evaluationen durchgeführt, die in den nachfolgenden Abschnitten beschrieben werden. Die technische Evaluation betrachtet die rein quantitative Richtigkeit der Antworten beider Systeme. Ziel dieser Evaluation ist es herauszufinden, welches der beiden Systeme die objektiv besseren Antworten auf Nutzer:in-Fragen liefert. Die empirische Evaluation betrachtet die Qualität der Antworten und deren Darstellungsform aus der Nutzer:in-Perspektive. Ziel dieser Evaluation ist es herauszufinden welches System, hinsichtlich der Zugänglichkeit und Nützlichkeit, den Nutzer:innen bessere Antworten liefert. Die Ergebnisse der Evaluationen werden in Abschnitt 5 beschrieben.

4.1 Technische Evaluation

Die technische Evaluation unterteilt sich in zwei Abschnitte. Im ersten Abschnitt wird die Verwendung von drei verschiedenen Übersetzern mit dem BERT-System verglichen. Dieser Vergleich ist notwendig, da für das BERT-System ein englisches Sprachmodell genutzt wird und das System auf Deutsch agieren soll. Im zweiten Abschnitt werden die beiden Systeme untereinander verglichen. Dabei verwendet das BERT-System dann den Übersetzer, der im ersten Teil am besten abgeschnitten hat.

Um beide Vergleiche der technischen Evaluation durchzuführen, wurde zunächst ein Fragenkatalog mit 15 verschiedenen Versicherungsszenarien erstellt. Die Teilnehmer:innen bekamen pro Versicherungsszenario die Aufgabe eine bestimmte Information mittels einer von ihnen formulierten Frage herauszufinden. Zur Aufstellung der Versicherungsszenarien wurden die „Allgemeinen Hausrat-Versicherungsbedingungen“ der Feuersozietät von 2009 verwendet. Die Versicherungsbedingungen (VB) unterteilen sich in drei unterschiedlich lange Abschnitte, die wiederum aus sich voneinander thematisch abgrenzenden Paragraphen mit Unterpunkten bestehen. Um eine willkürliche Auswahl der Themengebiete für den Fragebogen zu vermeiden, wurden aus den Abschnitten anteilig und zweistufig-randomisiert 15 Unterpunkte der einzelnen Paragraphen gezogen. Zu den 15 Unterpunkten wurden anschließend thematisch passende Versicherungsszenarien erstellt. Insgesamt beantworteten 18 Teilnehmer:innen den Fragebogen und somit ergab sich vor Bereinigung der Daten ein Fragenkatalog von 270 Fragen. Nach Bereinigung der gesammelten Fragen beinhaltete der Fragenkatalog 248 Fragen.

Um den Einfluss der Übersetzung der VB zu vergleichen, wurden diese mit drei verschiedenen Online-Übersetzern übersetzt. Hierfür wurden die Übersetzer von DeepL[1], Google[2] und Amazon[3] genutzt. Als Kontext wurden die übersetzten Versicherungsbedingungen dem BERT-System übergeben. Die Fragen aus dem Fragenkatalog wurden dem BERT-System in schriftlicher Form übergeben. Die schriftliche Form wurde hier bewusst gewählt, da somit der Einfluss einer unterschiedlichen Aussprache auf das Ergebnis vermieden wurde. Bei der Auswertung wurden die Antworten des Systems auf fachliche Richtigkeit geprüft. Eine Antwort wurde als richtig gewertet, wenn sich der wiedergegebene Text im richtigen Unterpunkt der VB befand. Eine solche richtige Antwort wurde mit einem Punkt bewertet. Ein halber Punkt wurde vergeben, wenn sich die Antwort im thematisch passenden Paragraphen der VB befand. Alle Antworten außerhalb des Paragraphen wurden mit null Punkten bewertet. Das BERT-System gibt mehrere Antworten, sortiert nach absteigender Wahrscheinlichkeit der Korrektheit, zurück. Das wurde berücksichtigt, indem die vergebenen Punkte mit dem Rang gewichtet wurden. Somit wurde eine richtige Antwort an erster Stelle höher bewertet als eine Antwort mit einem niedrigeren Rang.

Beim SCIO-System muss keine Gewichtung mit dem Rang vorgenommen, da das System nur eine Antwort zurück gibt. Zudem wurden die Antworten des SCIO-Systems nur mit null oder einem Punkt bewertet, da das System in der Regel nur eindeutig falsche bzw. richtige Antworten zurückgibt. Die Fragen wurden dem SCIO-System wie dem BERT-System in schriftlicher Form übergeben.

4.2 Empirische Evaluation

Für die empirische Evaluation wurde als Evaluationsinstrument eine Online-Umfrage gewählt. In der Umfrage wurde die Interaktion mit den Systemen anhand von beschreibenden Texten und Abbildungen der Interaktionsschnittstelle dargestellt. Es konnte kein Laborexperiment durchgeführt werden, da die Evaluation während der Corona-Pandemie durchgeführt wurde. Unter anderen Umständen wäre ein Laborexperiment die präferierte Methode gewesen.

Die Online-Umfrage ist folgendermaßen aufgebaut: Im ersten Abschnitt wird der demografische Datenwert Alter abgefragt und die Teilnehmer:innen sollen anhand einer 5-Punkte-Likert Skala mit drei Items ihre Versicherungskenntnisse einschätzen. Im Anschluss wird randomisiert ausgewählt, welches System den Teilnehmer:innen als Erstes gezeigt wird. Jedes System hat sechs simulierte Interaktionen über Versicherungsgegenstände wie z. B. Sturm oder Rohrbruch. Zur Repräsentation eines Systems wurde den Teilnehmer:innen eine mögliche gestellte Frage

[1] `https://www.deepl.com/de/translator`
[2] `https://translate.google.com/?sl=de&tl=en&op=docs`
[3] `https://aws.amazon.com/de/translate/`

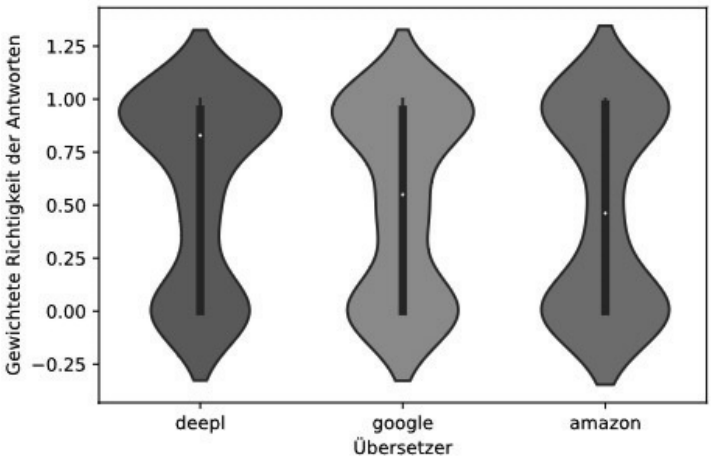

Abbildung 1 – Verteilung der Antworten der drei Übersetzer

eines/einer Nutzer:in und die Ausgabe des Systems gezeigt. Die Ausgabe des Systems wurde über eine Abbildung der Bildschirmoberfläche des Systems repräsentiert. Bei dem SCIO-System wurde zusätzlich ein separater Text angezeigt, der bei einer direkten Interaktion mit dem System als Audio ausgegeben wird. Die Ausgaben der Systeme beinhalten nur den „Best-Case“ einer Antwort des jeweiligen Systems, das heißt nur die richtige Antwort wird (als erstes Ergebnis) angezeigt. Jede simulierte Interaktion wird über eine 5-Punkte-Likert-Skala mit vier Items bewertet: „Ich bin überzeugt davon alle relevanten Informationen bekommen zu haben.“ (AB1), „Die Frage wurde angemessen ausführlich beantwortet“ (AB2), „Mir ist es leicht gefallen die Antwort auf die Frage zu finden“ (AB3), „Ich bin mit der sprachlichen Qualität der Antwort zufrieden“ (AB4). Nach der Bewertung aller einzelnen Interaktionen des Systems wurde das gesamte System über eine Kurzversion des User Experience Questionnaire (UEQ) [4] bewertet. Von insgesamt 26 teilnehmenden Personen haben 18 die Online-Umfrage vollständig und verwertbar ausgefüllt.

Für die Auswertung der Items, die über eine 5-Punkte-Likert-Skala bewertet wurden, wurden die Skalenwerte auf die Werte von -2 („Trifft nicht zu“) bis 2 („Trifft zu“) abgebildet. Im Anschluss wurde mit den skalierten Werten der arithmetische Mittelwert und das Konfidenzintervall für ein Signifikanzniveau von $\alpha = 0,05$ für verschiedene Betrachtungsweisen der Items berechnet. Die Auswertung der UEQ-Items erfolgte über ein speziell für den UEQ konzipierten Datenanalyse-Tool[4]. Die Angaben auf dem UEQ werden auf Werte von -3 (negative Bewertung) bis 3 (positive Bewertung) abgebildet.

5 Ergebnisse

5.1 Selektion der Systemantworten

Abbildung 1 stellt die Bewertung der Antworten unter Berücksichtigung der Rangordnung aller Übersetzer dar. In dem Violin-Plot kann man die Verteilung der Daten und die Stärke der Verteilung in den einzelnen Bereichen ablesen. Bei allen Übersetzern lässt sich im oberen Bereich die stärkste Ausprägung erkennen. Besonders deutlich ist die Ausprägung bei dem Übersetzer von DeepL. Die geringste Ausprägung ist bei dem Übersetzer von Amazon zu erkennen. Bei

[4] `https://www.ueq-online.org`

Tabelle 1 – Arithmetische Mittelwerte der Richtigkeit für alle Antworten

Übersetzer	Arithm. Mittelwert	Standardabweichung
DeepL	0,59	0,43
Google	0,53	0,43
Amazon	0,49	0,45

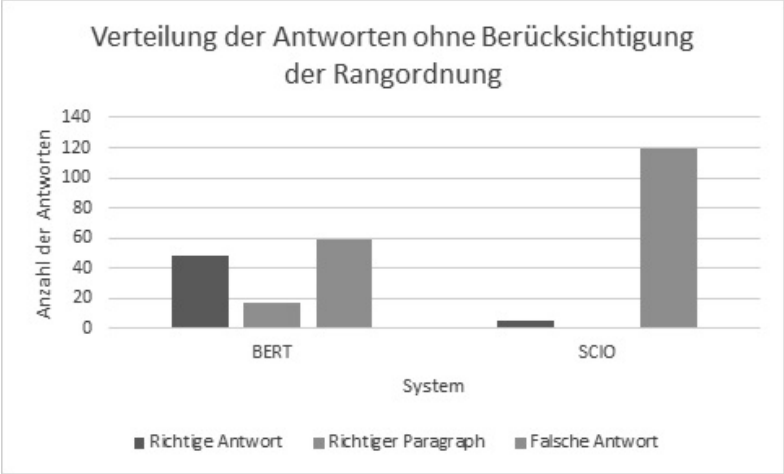

Abbildung 2 – Verteilung der Antworten beider Systeme ohne Berücksichtigung der Rangordnung

allen Übersetzern ist eine Abnahme der Ausprägung im mittleren Bereich und eine Zunahme im unteren Bereich zu erkennen. Die stärkste Ausprägung im unteren Bereich weist der Übersetzer von Amazon vor und die geringste der Übersetzer von DeepL.

Tabelle 1 enthält den arithmetischen Mittelwert und die Standardabweichung aller Antworten unter Berücksichtigung der Rangordnung für alle Übersetzer. Den höchsten arithmetischen Mittelwert aller Antworten weist der Übersetzer von DeepL mit 0,59 vor und den geringsten mit 0,49 der Übersetzer von Amazon.

In Abbildung 2 ist die Verteilung der Antworten ohne Berücksichtigung der Rangordnung der beiden Systeme zu sehen. Das BERT-System gibt insgesamt 48 richtig Antworten zurück, findet 17 richtige Paragraphen und beantwortet 59 Fragen falsch. Das SCIO-System gibt fünf richtige Antworten und 119 falsche Antworten zurück. Es findet keinen richtigen Paragraphen, da diese Kategorie nur vom BERT-System erfüllt werden kann. Somit gibt das BERT-System insgesamt mehr richtige Antworten zurück und die wenigsten falschen Antworten.

Tabelle 2 enthält den arithmetischen Mittelwert und Standardabweichung beider Systeme für alle Antworten. Bei dem BERT-System wird die Rangordnung der Antworten miteinbezogen. Der arithmetische Mittelwert für das BERT-System ist mit 0,42 größer als der Mittelwert des SCIO-Systems mit 0,04.

Tabelle 2 – Arithmetische Mittelwerte der Richtigkeit für alle Antworten beider Systeme unter Berücksichtigung der Rangordnung

System	Arithm. Mittelwert	Standardabweichung
BERT	0,42	0,43
S-CIO	0,04	0,20

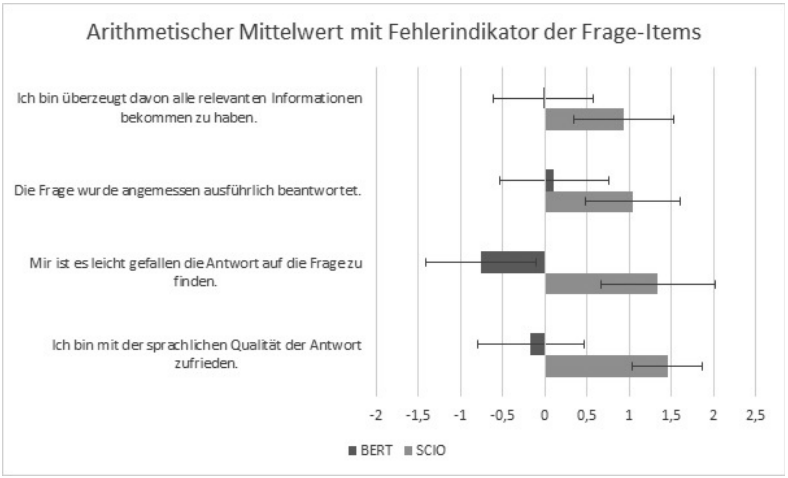

Abbildung 3 – Ergebnisse zur Bewertung der einzelnen Items zu allen Fragen für beide Systeme

In Abbildung 3 werden die arithmetischen Mittelwerte und zugehörigen Konfidenzintervalle der einzelnen Items AB1-AB4 abgebildet. Der arithmetische Mittelwert, sowie die Konfidenzintervalle befinden sich beim SCIO-System alle im positiven Bereich. Den höchsten arithmetischen Mittelwert weist das Item AB4 vor und den geringsten das Item AB1. Bei dem BERT-System befinden sich alle arithmetischen Mittelwerte im negativen Bereich bis auf das Item AB2. Damit weist dieses Item den höchsten arithmetischen Mittelwert vor und das Item AB3 den niedrigsten. Die Konfidenzintervalle beider Systeme überschneiden sich bei den Items AB1 und AB2. Die Anwendung des nichtparametrischen statistischen Wilcoxon-Tests ergab bei allen vier Items einen p-Wert der kleiner als 0,05 ist.

Bei der Bewertung über den UEQ zeigt sich, dass das SCIO-System bei der hedonischen und pragmatischen Qualität besser bei den befragten Personen abschneidet (siehe Abbildung 4). Bei beiden Qualitätsarten weist das SCIO-System eine statistisch signifikant bessere Bewertung bei einem Level von 5 % vor. Die hedonische Qualität schneidet bei beiden Systemen im Vergleich zur pragmatischen schlechter ab.

5.2 Bewertung der Systemantworten

Bei der Betrachtung des Violin-Plots und des arithmetischen Mittelwerts zeigt sich, dass der Übersetzer von DeepL am besten geeignet ist. Der Violin-Plot von DeepL ist im oberen Bereich am weitesten ausgeprägt und hat den höchsten arithmetischen Mittelwert, daher liefert das System mit diesem Übersetzer am meisten richtige Antworten, die zusätzlich einen hohen Rang haben. Es zeigt sich, dass für alle weiteren Vergleiche der Übersetzer von DeepL in Kombination mit dem BERT-System genutzt werden sollte.

Beim Vergleich der Antworten beider Systeme schneidet das BERT-System besser als das SCIO-System ab. Das BERT-System hat in Summe 51,5 mehr richtige Antworten (und Paragraphen) zurückgegeben als das SCIO-System, wenn der Rang nicht miteinbezogen wurde. Wird die Rangordnung mit einbezogen, schneidet das BERT-System insgesamt auch deutlich besser ab als das SCIO-System. Somit schneidet das BERT-System in Hinblick auf die technische Evaluation besser ab und liefert quantitativ mehr richtige Antworten als das SCIO-System.

Betrachtet man die Bewertung der einzelnen Items AB1-AB4 (s. Abbildung 3), zeigt sich, dass das BERT-System bei jedem Item schlechter bewertet wurde als das SCIO-System. Insbesondere die Konfidenzintervalle der Items AB3 und AB4 zeigen, dass das SCIO-System hier

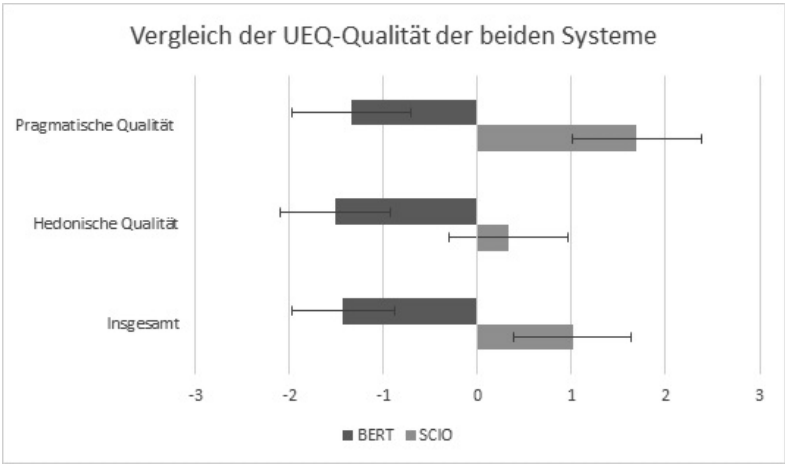

Abbildung 4 – Ergebnisse zur Bewertung der pragmatischen, hedonischen und gesamten Qualität für beide Systeme

deutlich besser abgeschnitten hat. Die Items geben an, wie leicht man die richtige Antwort finden kann und wie zufrieden die befragten Personen mit der sprachlichen Qualität der Antworten sind. Damit zeigt sich, dass das BERT-System in diesen beiden Kategorien besonders große Defizite im Vergleich zum SCIO-System besitzt. Durch die statistische Signifikanz des Wilcoxon-Tests kann eine Unterscheidung der zentralen Tendenzen der beiden Systeme angenommen werden.

Die Bewertung des UEQ zeigt, dass das SCIO-System hinsichtlich der pragmatischen und hedonischen Qualität besser wahrgenommen wird. Wie Abbildung 4 zeigt wird insbesondere die pragmatischen Qualität als hochwertiger wahrgenommen. Mittels eines Wilcoxon-Test kann eine Unterscheidung der zentralen Tendenzen der beiden Systeme hinsichtlich der pragmatischen und hedonischen Qualität gezeigt ($p < 0.05$) werden werden.

6 Diskussion

Die technische Evaluation zeigt, dass das sprachmodellbasierte BERT-System quantitativ die besseren Ergebnisse als das regelbasierte System liefert. Das BERT-System konnte in einigen Themenbereichen auf unterschiedlich formulierte Fragen die gleiche richtige Antwort finden. Diese Eigenschaft bietet einen Vorteil bei der Interaktion mit Nutzer:innen. Das SCIO-System hat hingegen Schwierigkeiten mit Fragen, die nicht mit dem gleichen oder ähnlichen Wortlaut in der Datenbank vorhanden sind. Um die Ergebnisse des SCIO-Systems hinsichtlich der Quantität der richtigen Antworten zu verbessern, sollte die Datenbank mit weiteren Themenbereichen aus den Versicherungsbedingungen und Formulierungen der Fragen erweitert werden. Das bessere Abschneiden des BERT-Systems könnte durch die Ausgabe von mehreren Antworten gesteigert worden sein. Bewertet man das BERT-System mit strengeren Kriterien, z. B. wenn nur Antworten als richtig anerkannt werden, die zusätzlich noch auf Rang eins erscheinen, würden nur 12 Antworten als richtig gelten. Zwar sind das immer noch mehr richtige Antworten als das SCIO-System, aber der Abstand ist deutlich geringer. Jedoch könnte dieser Wert bei dem BERT-System noch verbessert werden, wenn das Sprachmodell zum Beispiel speziell auf Frage-Antwort-Paare aus dem Versicherungskontext trainiert wird. Möglicherweise würde auch ein Training des deutschen BERT-Modells mit Frage-Antwort-Paaren eine Verbesserung erzielen, da dadurch Fehler durch die Übersetzung ausgeschlossen werden können. Außerdem könnte das BERT-System mit weiteren Modellen kombiniert werden, die möglicherweise besser für die

Aufgabe geeignet sind. Beispielsweise gibt es ein weiteres Transformersmodell, das sich speziell für die Verarbeitung von längeren Texten eignet [5].

Die empirische Evaluation zeigt, dass das regelbasierte SCIO-System von den befragten Personen als besser im Vergleich zum BERT-System wahrgenommen wird. Ein Grund dafür könnte die Übersetzung der Antworttexte sein, da Nutzer:innen mit der Qualität der Sprache deutlich unzufriedener waren (siehe Abbildung 3). Nutzer:innen ist es mit dem BERT-System schwergefallen die richtige Antwort zu finden. Das könnte an der Wiedergabe von mehreren Antwortmöglichkeiten und größeren Textblöcken liegen, wodurch eine eher unübersichtliche Darstellungsform der Antworten entsteht. Zusätzlich wurden die Textblöcke aus den Versicherungsbedingungen wiedergegeben, die möglicherweise für Laien schwerer verständlich sein können. Das SCIO-System hat eine insgesamt übersichtlichere Darstellung der Antworten, die im Allgemeinen etwas kürzer und leichter formuliert sind. Die Auswertung des UEQ bestätigt die Ergebnisse der Items AB1-AB4. Durch die eben genannten Gründe könnte daher auch die pragmatische und hedonische Qualität beim SCIO-System besser sein. Das im Allgemeinen bessere Abschneiden des SCIO-Systems könnte daran liegen, dass die Technik, auf die das System aufbaut, speziell für die Sprachinteraktion mit Nutzer:innen konzipiert wurde. Das Sprachmodell des BERT-Systems ist nicht explizit für eine Sprachinteraktion designt. Das BERT-System müsste an eine Sprachinteraktion mit Nutzer:innen angepasst werden, um das Erlebnis einer Sprachinteraktion mit dem System zu verbessern.

Abschließend zeigt sich, dass beide Systeme zum jetzigen Zeitpunkt noch keinen optimalen Zustand erreicht haben und weitere Arbeit nötig ist, um die Systeme zu verbessern. Da die Stärken und Schwächen der beiden Systeme konträr sind, sollte untersucht werden, inwiefern eine Orchestrierung beider System möglich ist und ob diese zu einer Verbesserung der Quantität und Qualität führen kann.

Literatur

[1] DEVLIN, J., M. CHANG, K. LEE, und K. TOUTANOVA: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`. `1810.04805`.

[2] HUGGINGFACE: *DistilBERT*. Huggingface. Zugriff auf: `https://huggingface.co/transformers/model_doc/distilbert.html`, o.J. (20.01.2022).

[3] RAJPURKAR, P., J. ZHANG, K. LOPYREV, und P. LIANG: *Squad: 100,000+ Questions for Machine Comprehension of Text*. 2016. `1606.05250`.

[4] SCHREPP, M., A. HINDERKS, und J. THOMASCHEWSKI: *Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). International Journal of Interactive Multimedia and Artificial Intelligence*, 4, S. 103, 2017. doi:10.9781/ijimai.2017.09.001.

[5] BELTAGY, I., M. E. PETERS, und A. COHAN: *Longformer: The Long-Document Transformer*. 2020. `2004.05150`.