OTH REGENSBURG

Sprachassistenten

Anwendungen, Implikationen, Entwicklungen

2. ITG-Workshop, Regensburg, 4. März 2024

Herausgeber: Timo Baumann und Ingo Siegert

VDE ITG
Informationstechnische Gesellschaft

Herausgeber OTH Regensburg, Timo Baumann, Ingo Siegert

Inhaltsverzeichnis

Vorwort **1**

Eingeladene Vorträge **3**

Sebastian Merkel
Smarte Lautsprecher in Gesundheit und Pflege – Anwendungsfälle und Motivationen. Ergebnisse eines Scoping Reviews 4

Carolin Wienrich
Verstehen und Verstanden werden: Psychologische Aspekte der sprachbasierten Mensch-KI Interaktion 5

Andreas M. Klein
Towards context-dependent UX measurement for Voice User Interfaces (VUIs) 6

Steffen Werner
HeyMercedes Voice Assistant: Shaping the future of in car user experience . . 7

Kurzfassungen der Beiträge **9**

Zhengyang Li, Timo Lohrenz, Matthias Dunkelberg, Tim Fingscheidt
Real-Time Interactive Demonstrator for Audiovisual Speech Recognition and Lip Reading 9

Anna Leschanowsky, Birgit Popp, Nils Peters
Debiasing-Strategien für mehr Datenschutz und Sicherheit in Sprachassistenzsystemen 12

Jan Nehring
Modulare Dialogsysteme 14

Lia Frischholz, Lisa Winkler, Christian Gaida, Melanie Schindler, Rico Petrick, Felix Gräßer
Entwicklung eines Voicebot-Frameworks am Beispiel Verkehrsauskunft für einen Nahverkehrsbetrieb 16

Stefan Hillmann, Philipp Harnisch
MIA-PROM Sprachassistent zur Erhebung von PROM-Fragebögen in der Reha 19

Matthias Busch, Long Nguyen, Ingo Siegert
Voice Interaction in Motion: Eaasy VUI and Physical Exertion 22

Martha Schubert, Michael Schenk, Julia Krüger, Melanie Elgner, Florian Junne, Ingo Siegert
First Steps into ASPIRE: A Pilot Study on Automated Speech Analysis Regarding Psychotherapeutic Alliance in Psychotherapies 24

Stefan Schaffer, Aaron Ruß
Konversationelle Interaktionen für Hybride Veranstaltungen 26
Oliver Jokisch, Karl M. Walter
Potentials of Chatbots in the German Public Administration 28

Vorwort

Der zweite ITG-Workshop „Sprachassistenten – Anwendungen, Implikationen, Entwicklungen“ findet am 5. März 2024 in Regensburg statt. Er bietet eine organisatorische und inhaltliche Fortführung des ersten Workshops vor vier Jahren in Magdeburg 2020. Auch in diesem Jahr ist er wieder der Konferenz Elektronische Sprachsignalverarbeitung angegliedert.

Die Entwicklung im Bereich digitaler Sprachassistenzsysteme hat sich in den vergangenen Jahren weiter beschleunigt. Apple’s Siri, Amazon’s Alexa oder der Google Assistant sind im Alltag von immer mehr Menschen angekommen. Ihre Attraktivität ist in der einfachen Bedienung begründet, sie gestatten es uns Internetrecherchen, Online-Bestellungen, Raumüberwachungen und andere Smart-Home-Dienste nur durch Zuruf durchzuführen.

Die hohe Natürlichkeit der sprachbasierten Interaktion verschleiern der Nutzerin oder dem Nutzer mitunter die Einschränkungen der Applikationen, insbesondere, da heutige Sprachassistenten teilweise nur bessere Fernbedienungen sind und das Gerät nur über einen Teil der Funktionen sprachliches Feedback gibt. Zukünftig sollen diese jedoch nicht nur einfache Befehle verarbeiten, sondern auch eine natürliche und reibungslose Interaktion ermöglichen. Hierzu ist neben technischen Verbesserungen der Spracherkennung auch eine verbesserten Sprachverständnis sowie intelligentere Dialogführung nötig. Weiterhin gehören neben neuen technischen Lösungen aber auch rechtliche Aspekte, die sich durch die Verbreitung von Sprachassistenten und der strengeren Datenschutzregelungen ergeben. Wann und was mitgehört wird ist häufig nicht ersichtlich und auch was mit den Sprachdaten geschieht weiß der Nutzer zumeist nicht. Dies kann zu Akzeptanzproblemen führen.

Auf dem Workshop werden vielfältige und interdisziplinäre Beiträge in eingeladenen Vorträgen und als eingereichte Poster präsentiert. Durch die gute Mischung von Beitragenden sowohl aus der Hochschullandschaft als auch aus der Industrie werden die verschiedensten Aspekte anwendungsnah diskutiert.

Die Organisatoren des Workshops danken dem ITG Fachbereich „Dienste und Anwendungen“ für die Unterstützung, der OTH Regensburg und den Sponsoren der ESSV, *alphaspeech* und *Genie Enterprises* für die finanzielle Unterstützung, sowie allen Autoren und Teilnehmern für ihre aktive Teilnahme, die diesen Workshop erst zu einem erfolgreichen Event machen. Wir freuen uns bereits jetzt auf eine Neuauflage in den kommenden Jahren.

Regensburg, 5. März 2024 Timo Baumann, Ingo Siegert

Eingeladene Vorträge

Smarte Lautsprecher in Gesundheit und Pflege – Anwendungsfälle und Motivationen. Ergebnisse eines Scoping Reviews

Sebastian Merkel
Ruhr-Universität Bochum

Der Einsatz von Spracherkennung und -verarbeitung im Gesundheits- und Pflegesektor hat bereits eine lange Tradition, hat vor dem Hintergrund neuer technischer Entwicklungen aktuell einen neuen Schub erhalten. Im Rahmen eines Scoping Reviews wurde der Einsatz sog. Smarter Lautsprecher in Gesundheit und Pflege untersucht. Im Mittelpunkt steht vor allem die Frage, welche Anwendungsfälle sich identifizieren lassen und welche Vorteile Entwickler:innen in dieser Technologie sehen bzw. welche Versorgungsbedarfe identifiziert wurden, die mittels der Technologie adressiert werden sollen.

Sebastian Merkel ist Inhaber der Juniorprofessur für Gesundheit und E-Health an der Fakultät für Sozialwissenschaft der Ruhr-Universität Bochum und zuvor Forschungsdirektor des Schwerpunktes „Gesundheitswirtschaft und Lebensqualität" am Institut Arbeit und Technik der Westfälischen Hochschule. Studium der Politikwissenschaft, Geschichte, Sozialwissenschaft und Politikmanagement in Bochum und Duisburg-Essen, Promotion an der Universität Witten/Herdecke. Die Juniorprofessur erforscht die Auswirkung der Digitalisierung auf den Gesundheitssektor. Im Mittelpunkt stehen dabei Fragen danach, wie (digitale) Technik partizipativ entwickelt und gestaltet werden kann und welche Faktoren sich auf die Implementation digitaler Technik wie auswirken.

Verstehen und Verstanden werden: Psychologische Aspekte der sprachbasierten Mensch-KI Interaktion

Carolin Wienrich
Universität Würzburg

Interaktionen mit sprachbasierten KI-Systemen sind intuitiv und knüpfen durch Dialogfähigkeiten an Interaktionsgewohnheiten an, die bisher zwischen Menschen stattgefunden hat. Der Vortrag zeigt, wie diese Interaktionsgewohnheiten auf sprachbasierte KI-Systeme übertragen werden und wie diese Übertragung die Interaktion beeinflusst. Darüber hinaus zeigt er auf, wie die Gestaltung der artifiziellen Interaktionspartner:innen, Nutzer:innen beeinflussen können und zeigt Potenziale und Risikien, die daraus erwachsen. Zudem stellt Carolin Wienrich einige interaktive Trainings und Evaluationen vor, die insbesondere das Wissen über sprachbasierte KI-Systeme verbessern.

Carolin Wienrich erforscht die Schnittstelle von Psychologie und Technologie. Dabei untersucht sie Menschen in der Interaktion mit künstlichen Entitäten und Gelingensbedingungen für eine menschzentrierte KI-Entwicklung. Sie entwickelt Messinstrumente und interaktive Trainings zum Aufbau von KI-Kompetenzen. Ein weiterer Schwerpunkt liegt in der Erforschung von menschlichen Verhalten sowie Potenzialen und Risiken von eXtended Realities. Sie studierte und promovierte in der Psychologie und ist aktuell mit ihrer Professur für Psychologie intelligenter interaktiver Systeme an der Universität Würzburg im Feld der Mensch-Computer Interaktion tätig. Außerdem sie leitet sie den XR-HUB Würzburg und ist Teil des KI-Zentrums CAIDAS in Würzburg.

Towards context-dependent UX measurement for Voice User Interfaces (VUIs)

Andreas M. Klein
University of Applied Sciences Emden/Leer, University of Seville, Spain

VUIs are trendy and highly available, but challenges such as privacy, understanding the dialogue, and its context remain. Therefore, we see a demand for UX assessment for VUIs considering the context of use. We identified and categorized 32 UX aspects for VUIs and constructed voice quality scales. Furthermore, we split the VUI context of use into quantifiable parts, which we mapped to relevant UX qualities. Our method considered contextual UX by combining the concepts of the context of use and UX. This approach helps to describe the VUI context of use for realizing the context-dependent UX measurement recommendations for VUIs.

Andreas M. Klein ist Lehrbeauftragter und wissenschaftlicher Mitarbeiter im Fachbereich Technik der Hochschule Emden/Leer, Deutschland. Er unterrichtet in den Online-Kursen Medien- und Wirtschaftsinformatik und im Präsenzkurs Medientechnik. Seit September 2020 ist Andreas Doktorand im Fachbereich Computersprachen und -systeme an der Universität Sevilla, Spanien. Der Forschungsschwerpunkt seiner Dissertation liegt auf der Untersuchung von Sprachassistenzsystemen. Andreas hat ein Diplom in Elektrotechnik und einen Master of Engineering in Technical Management.

HeyMercedes Voice Assistant: Shaping the future of in car user experience

Steffen Werner
Mercedes-Benz AG

Wie sieht die Zukunft der fahrzeuggebunden Sprachassistenten aus? Diese Frage wird am Beispiel des Mercedes Sprachassistenten diskutiert. Dazu wird kurz in die bisherige Evolution von HeyMercedes geschaut und wie sich diese im Vergleich zur mobilen Konkurrenz einordnet. Es wird betrachtet, welche Sprachdialoganteile werden besonders häufig (und welche besonders selten) verwendet werden und wie sich daraus datengetriebene Entwicklungsansätze ableiten lassen (oder auch nicht). Die Verbindung von fahrzeugspezifischen Daten mit innovativen Konversationsmöglichkeiten moderner AI-Systeme ergeben viele neue Möglichkeiten Spitzentechnologie nahtlos ins Fahrzeug zu integrieren. Es werden erste Ergebnisse einer Pilot-Integration von ChatGPT in das Mercedes-Dialogsystem vorgestellt. Diese Umsetzung zeigt die erweiterten Möglichkeiten in Verbindung mit fahrzeugspezifischen Assistenten und setzt die Richtung für erweiterte Bedienmöglichkeiten über alle Sprachen und Märkte hinweg.

Steffen Werner bringt als Projektleiter für Sprach-Assistenzsysteme der Mercedes-Benz AG die neueste Sprachtechnologie in die Fahrzeuge und damit auf die Straße. Mit einem interkulturellen und global verteilten Projektteam fokussiert sich Steffen auf die softwaretechnische Integration von modernster Sprach-Technologie, welche nach erfolgreicher Evaluierung und Erprobung in die weltweite Serienproduktion überführt wird. Steffen konzentriert sich dabei auf einen hybriden Systemansatz, welcher die Vorteile von Off- und Online-Software-Lösungen verbindet und dem Fahrer eine optimal abgestimmte User-Experience für die verschiedenen Fahr- und Bedien-Modi bereitstellt.

Kurzfassungen der Beiträge

REAL-TIME INTERACTIVE DEMONSTRATOR FOR AUDIOVISUAL SPEECH RECOGNITION AND LIP READING

Zhengyang Li, Timo Lohrenz, Matthias Dunkelberg, Tim Fingscheidt
Technische Universität Braunschweig,
Institute for Communications Technology,
30106 Braunschweig, Germany
{zhengyang.li, t.fingscheidt}@tu-bs.de

BACKGROUND: Particularly in noisy and multi-talker conditions, audiovisual speech recognition (AV-ASR) [1,2,3] excels conventional automatic speech recognition (ASR) [4,5,6] based solely on acoustic input. Besides an ASR engine, AV-ASR employs lip-reading [7], which is a technique that utilizes the speaker's lip movements to recognize spoken utterances without relying on acoustic information, which has potential applications for helping people who suffer from aphonia and supporting crime investigation.

METHOD: In this work, we first trained three all-attention based transformer models [8] on the public Lip Reading Sentences 3 (LRS3) dataset [9]: AV-ASR, lip-reading model, and the conventional acoustic ASR. The AV-ASR and the lip-reading model utilize the pre-trained state-of-the-art audiovisual hidden unit BERT (AV-HuBERT) [10], and the acoustic ASR is trained from scratch.

Second, we developed a real-time web-based demonstrator with an interactive interface for these three models. Users can speak to the demonstrator, and the transcribed text from these three models is displayed on the screen in real-time. The demonstrator captures the voice using a microphone and detects the lip-movement of the speaker using a web-camera. Face landmark detection and face alignment are integrated into this demonstrator to center the lip region in the input video. Both, face landmark and centered lip region, can be visualized on screen in real-time to help the user to understand the video pre-processing. In addition, a voice activity detection can be activated to enable a permanent continuous speech recognition.

RESULTS: In Figure 1, we evaluate the performance of the three models in our demonstrator, i.e., AV-ASR, lip-reading model, and the conventional acoustic ASR. The word error rates (WERs, lower is better) are measured on the LRS3 test set with babble noise in various signal-to-noise ratios (a lower SNR means a more noisy condition).

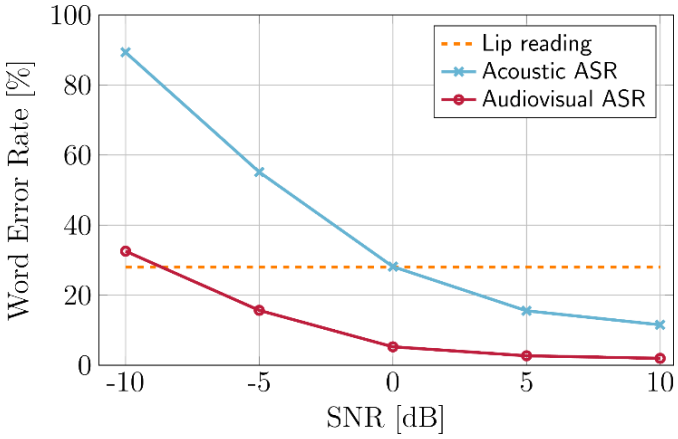

Figure 1 – WERs (%) of the AV-ASR, the lip-reading model, and the conventional acoustic ASR on LRS3 test set with babble noise in various SNRs.

The lip-reading model (seen in the orange curve) achieves a WER of 28.0% on the LRS3 test set. And the performance is not influenced by the noisy level, as the lip reading model leverage only visual

modalities. Expectedly, the performance of the AV-ASR and the lip-reading model improves with increasing SNRs. The AV-ASR (red curve) outperforms the acoustic ASR (blue curve) significantly in all SNRs. In addition, in more noisy conditions, the AV-ASR presents a larger WER improvement compared to acoustic ASR, showing the noise robustness of AV-ASR in noisy conditions.

CONCLUSIONS: In this work, we trained and evaluated three all-attention based transformer models on the LRS3 dataset, i.e., the audiovisual ASR, the lip-reading model, and the acoustic ASR. In addition, we developed a real-time web-based demonstrator for these three models. The demonstrator provides users a more engaging and informative way to explore the capabilities of these three models in different noise levels and understand their potential applications.

Funding acknowledgment: German BMWK, 01MK20011T, large-scale SPEAKER project.

LIST OF REFERENCES

[1] S. Receveur, R. Weiss, and T. Fingscheidt, "Turbo Automatic Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 846–862, May 2016.

[2] Z. Li, T. Lohrenz, M. Dunkelberg, and T. Fingscheidt, "Transformer-Based Lip-Reading with Regularized Dropout and Relaxed Attention," in *Proc. of SLT*, Doha, Qatar, Jan. 2023, pp. 723–730.

[3] Z. Li, T. Graave, J. Liu, T. Lohrenz, S. Kunzmann, and T. Fingscheidt, "Parameter-Efficient Cross-Language Transfer Learning for a Language-Modular Audiovisual Speech Recognition," in *Proc. of ASRU*, Taipei, Taiwan, Dec. 2023, pp. 1–8.

[4] T. Lohrenz, Z. Li, and T. Fingscheidt, "Multi-Encoder Learning and Stream Fusion for Transformer-Based End-to-End Automatic Speech Recognition," in *Proc. of Interspeech*, Brno, Czech Republic, Sep. 2021, pp. 2846–2850.

[5] T. Lohrenz, P. Schwarz, Z. Li, and T. Fingscheidt, "Relaxed Attention: A Simple Method to Boost Performance of End-to-End Automatic Speech Recognition," in *Proc. of ASRU*, Cartagena, Colombia, Dec. 2021, pp. 177–184.

[6] T. Lohrenz, B. Moller, Z. Li, and T. Fingscheidt, "Relaxed Attention for Transformer Models," in *Proc. of IJCNN*, Gold Coast, QLD, Australia, Jun. 2023, pp. 1–10.

[7] Z. Li, T. Lohrenz, M. Dunkelberg, and T. Fingscheidt, "Transformer-Based Lip-Reading with Regularized Dropout and Relaxed Attention," in *Proc. of SLT*, Doha, Qatar, Jan. 2023, pp. 723–730.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need," in *Proc. of NIPS*, Long Beach, CA, USA, Dec. 2017, pp. 1–11.

[9] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep Audio-Visual Speech Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–11, Dec. 2018, (early access).

[10] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Robust Self-Supervised Audio-Visual Speech Recognition," *arXiv:2201.02184*, Jul. 2022.

DEBIASING-STRATEGIEN FÜR MEHR DATENSCHUTZ UND SICHERHEIT IN SPRACHASSISTENZSYSTEMEN

Anna Leschanowsky[1]
[1]*Fraunhofer IIS*
anna.leschanowsky@iis.fraunhofer.de

HINTERGRUND: Mit der zunehmenden Verbreitung von Conversational AI (CAI)-Systemen, insbesondere Sprachassistenten, in Wohnungen, Autos und öffentlichen Räumen stehen Nutzende vor komplexen Entscheidungen bezüglich der Preisgabe persönlicher Informationen und der Verarbeitung ihrer Daten durch Anbietende. Gleichzeitig müssen Entwickelnde die Privatsphäre und Sicherheit der Nutzenden während der Systementwicklung berücksichtigen. Jedoch sind diese Bemühungen oft von kognitiven Verzerrungen beeinflusst. Debiasing- und Nudging-Strategien, die in verschiedenen Bereichen wie der Medizin, Luftfahrt und Politik, erfolgreiche Anwendung finden, bieten vielversprechende Ansätze, um Privatsphäre-schützende CAI-Systeme zu entwickeln. Unser Ziel ist es, Akteueren im CAI-Ökosystem etablierte Strategien an die Hand zu geben, um Privatsphäre und Sicherheit von CAI zu verbessern. Aufgrund der Besonderheiten von CAI, ist es nötig klassische Kategorisierungen zu kombinieren, um Orientierung und einen Überblick zu gewinnen. Daher stellen wir eine neuartige Kategorisierung von Debiasing-Strategien vor, zeigen, wie bestehende Strategien an die spezifischen Anforderungen von Sprachassistenzsystemen und CAI im Allgemeinen angepasst werden können und ordnen sie relevanten Akteuren im Sprachassistenz-Ökosystem zu [1].

METHODE: Basierend auf explorativer Literaturrecherche und bestehenden Kategorisierungen von Debiasing-Strategien [2], [3], etablieren wir eine neuartige Klassifizierung von Debiasing-Techniken für Conversational AI Systeme (siehe Abb. 1). Hierfür diskutieren wir zunächst Akteure des CAI-Ökosystems entsprechend rechtlicher Richtlinien [4] und identifizieren Hauptquellen für problematische Datenschutz- und Sicherheitsentscheidungen im CAI Kontext wie beispielsweise Informationsasymmetrie, Heuristiken und kognitive Verzerrungen [5], [6]. Zuletzt diskutieren wir existierende Debiasing-Techniken, passen diese an den Kontext von CAI an und ordnen sie den relevanten Stakeholdern im CAI-Ökosystem zu.

ERGEBNISSE: Nach Analyse bisheriger Debiasing-Frameworks schlagen wir eine zweidimensionale Kategorisierung für Debiasing-Strategien im Kontext von CAI vor (siehe Abb. 1). Hierbei bezieht sich die erste Dimension bezieht sich auf die zeitliche Komponente, nämlich das Erscheinen des Debiasing Effekts, während die zweite Dimension das Modifikationsobjekt, hier Person oder Umgebung, betrachtet. Die Berücksichtigung beider Dimensionen ist essentiell, da CAI-Systeme aufgrund natürlicher und nahtloser Interaktion mit Nutzenden sowohl die Umgebung als auch die Person modifizieren können mit Kurz- oder Langzeitfolgen. So können CAI-Systeme proaktiv als Mentoren oder Lehrer für Datenschutz und Sicherheit agieren und dadurch sowohl die Umgebung als auch die Einzelperson beeinflussen [1]. Des Weiteren betrachten wir Klassen nicht voneinander getrennt, sondern heben hervor, dass bestimmte Techniken wie kognitive Strategien sich in Zwischenbereichen befinden können. Unsere Analyse, Adaption und Zuordnung von Debiasing-Strategien zu relevanten Akteuren zeigt großes Potential für zukünftige Forschung in diesem Bereich. Insbesondere gilt es, bestehende und neue Debiasing-Strategien im Kontext von CAI zu entwickeln und zu evaluieren, die Auswirkungen unterschiedlicher Modalitäten (Stimme vs. Text) auf die Wirksamkeit der Strategien zu untersuchen und ein umfassendes Evaluationsframework für Debiasing-Strategien im Bereich Datenschutz, Sicherheit und Privatsphäre zu entwickeln.

SCHLUSSFOLGERUNGEN: Wir stellen einen ganzheitlichen Ansatz zur Kategorisierung von Debiasing-Strategien für CAI-Systeme vor, der sowohl zeitliche als auch typbasierte Perspektiven integriert. Unser Ziel ist es durch diesen Ansatz eine Grundlage für die Entwicklung von effektiven und legitimen Debiasing-Strategien für CAI-Systeme zu schaffen. Insbesondere soll unser Beitrag zur Diskussion anregen, gemeinsam die Entwicklung und Umsetzung dieser Strategien voranzutreiben.

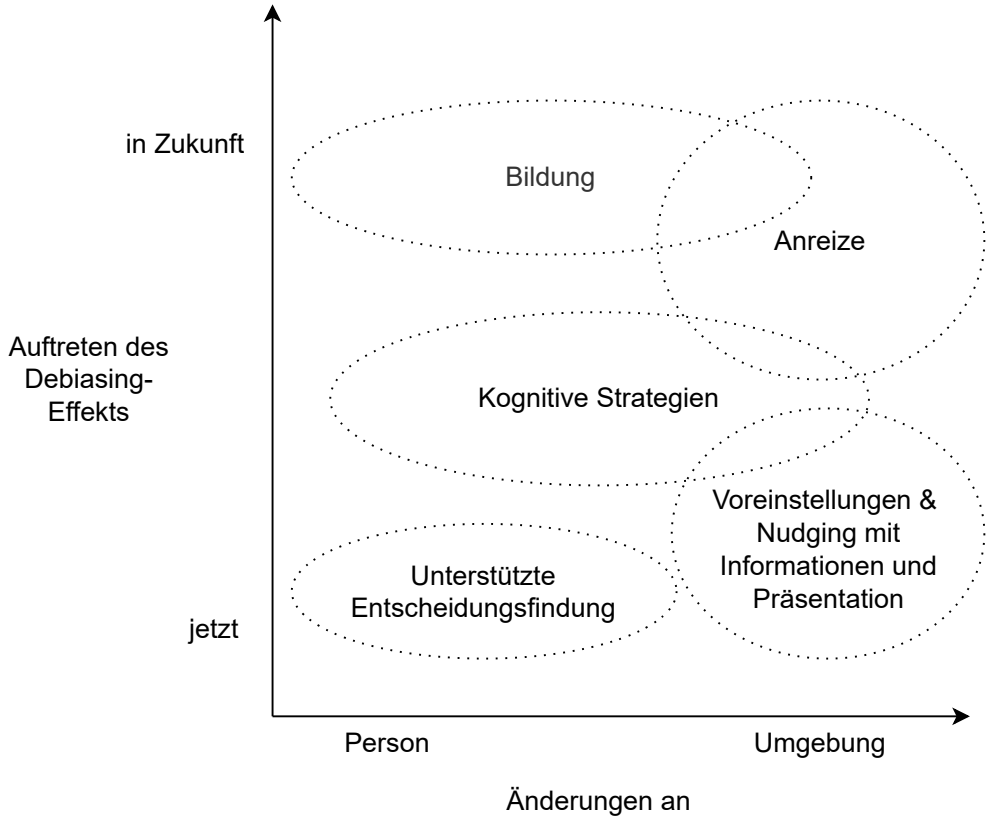

Abbildung 1 - Schematische Darstellung des Frameworks zur Kategorisierung von Debiasing-Strategien für Conversational AI [1].

LITERATURVERZEICHNIS

[1] A. Leschanowsky, B. Popp, and N. Peters, "Debiasing Strategies for Conversational AI: Improving Privacy and Security Decision-Making," *Digit. Soc.*, vol. 2, no. 3, p. 34, Sep. 2023, doi: 10.1007/s44206-023-00062-2.

[2] J. B. Soll, K. L. Milkman, and J. W. Payne, "A User's Guide to Debiasing," in *The Wiley Blackwell Handbook of Judgment and Decision Making*, G. Keren and G. Wu, Eds., Chichester, UK: John Wiley & Sons, Ltd, 2015, pp. 924–951. doi: 10.1002/9781118468333.ch33.

[3] P. Croskerry, G. Singhal, and S. Mamede, "Cognitive debiasing 2: impediments to and strategies for change," *BMJ Qual. Saf.*, vol. 22, no. Suppl 2, pp. ii65–ii72, Oct. 2013, doi: 10.1136/bmjqs-2012-001713.

[4] "Guidelines 02/2021 on virtual voice assistants | European Data Protection Board." Accessed: Feb. 02, 2024. [Online]. Available: https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-022021-virtual-voice-assistants_en

[5] A. Acquisti *et al.*, "Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online," *ACM Comput. Surv.*, vol. 50, no. 3, pp. 1–41, May 2018, doi: 10.1145/3054926.

[6] A. Acquisti, L. Brandimarte, and G. Loewenstein, "Privacy and human behavior in the age of information," *Science*, vol. 347, no. 6221, pp. 509–514, Jan. 2015, doi: 10.1126/science.aaa1465.

MODULARE DIALOGSYSTEME

Jan Nehring

Deutsches Forschungszentrum für Künstliche Intelligenz
jan.nehring@dfki.de

Dieser Abstract ist eine Zusammenfassung der Promotion des Autors, die am 6.12.2023 eingereicht wurde.

In der Praxis bestehen Dialogsysteme (DS) häufig aus mehreren Sub-DS, im Folgenden Agenten. Gründe dafür sind 1) dass ein Unternehmen mehrere Agenten implementiert hat und diese nun ohne den Aufwand einer Neuimplementierung kombinieren möchte. 2) kann ein solcher föderaler Aufbau eine Anforderung sein, z.B. aus Gründen der Softwarearchitektur. Ein dritter Grund 3) ist die Kombination von sonst inkompatiblen Technologien. Wir nennen diese Architektur "Modular Dialog Systems" (MDS). Die Forschung an solchen Kombinationen von Dialogsystemen ist eine Lücke in der wissenschaftlichen Literatur. Obwohl zahlreiche Arbeiten mehrere Dialogagenten miteinander kombinieren, sind generelle Untersuchungen über MDS so gut wie nicht vorhanden.

Die Promotion untersucht drei Forschungsfragen (FF): FF1) Charakteristiken MDS, FF2) Techniken für die Zuordnung einer eingehenden Benutzeräußerung zu einem der Agenten (die s.g. Module Selection (MS)) und FF3) wie die Kombination mehrerer Agenten teuren GPU-Arbeitsspeicher einsparen kann. Dazu werden drei Experimente durchgeführt.

- Experiment eins kombiniert mehrere aufgabengetriebene Agenten zu einem MDS und untersucht FF1 und 2. Der Aufbau des Experiments ist einem früherem Experiment ähnlich [1, 2] und macht die Ergebnisse mehrerer vorhergehender Publikationen [2, 3, 4, 5] vergleichbar.
- Experiment zwei kombiniert aufgabengetriebenen Dialog mit Question Answering für FF1 und wird in Nehring et al. [6] näher beschrieben.
- Experiment drei untersucht für FF3 den Einsatz von Adaptern [7, 8], um beim parallelen Deployment mehrerer NLU Systeme GPU- Arbeitsspeicher einzusparen. Es wird in Nehring et al. [9] näher beschrieben.

Es werden verschiedene Evaluationsmaße für MDS vorgeschlagen. Zum einen wird der Unterschied der Performance eines MDS verglichen mit den gleichen Funktionen implementiert als ein einzelnes System. Zum anderen wird die Qualität der MS gemessen. Außerdem werden zahlreiche Ergebnisse vorgestellt. Die Auswirkung verschiedener MDS-Charakteristika wie Anzahl von Agenten und Domänen auf die Performance werden untersucht. Die Performance von MDS ist meist niedriger als die seines nicht-modularen Gegenstücks. In manchen Fällen ist sie jedoch höher, nämlich wenn eine starke MS auf schwache NLU-Systeme trifft. Verschiedene überwachte und nicht-überwachte Ansätze für MS, unter anderem ein neues Modell, werden vorgestellt. Es wird gezeigt, dass Adapter GPU-Arbeitsspeicher einsparen können beim parallelen Deployment von NLUs.

Literatur

[1] BERK, R. M.: *Context Aware Module Selection in Modular Dialogue Systems*. Master's thesis, TU Berlin, Straße des 17. Juni 135, 10623 Berlin, 2022.

[2] NEHRING, J., R. M. BERK, und S. HILLMANN: *Context aware module selection in modular dialogue systems*. In *Proceedings of RANLP2023*. 2023.

[3] NEHRING, J., A. AHMED, und L. A. JAGER: *Module selection: A new task for dialog systems*. In *Proceedings of the 13th International Workshop on Spoken Dialogue Systems Technology*, IWSDS'23. 2023. URL `https://drive.google.com/file/d/1ndI1ZuXV6oXcrF0IGcv-s9doq0-sfm0-/view?usp=sharing`.

[4] NEHRING, J. und A. AHMED: *Normalisierungsmethoden für Intent Erkennung Modularer Dialogsysteme*. In S. HILLMANN, B. WEISS, T. MICHAEL, und S. MÖLLER (Hrsg.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2021*, S. 264–271. TUDpress, Dresden, 2021. URL `https://essv.de/essv2021/pdfs/18_nehring.pdf`.

[5] GÖRZIG, P., J. NEHRING, S. HILLMANN, und S. MÖLLER: *A comparison of module selection strategies for modular dialog systems*. In C. DRAXLER (Hrsg.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2023*, S. 40–47. TUDpress, Dresden, 2023. URL `https://www.essv.de/pdf/pdf/2023_40_47.pdf`.

[6] NEHRING, J., N. FELDHUS, H. KAUR, und A. AHMED: *Combining open domain question answering with a task-oriented dialog system*. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, S. 38–45. Association for Computational Linguistics, Online, 2021. URL `https://aclanthology.org/2021.dialdoc-1.5`.

[7] HOULSBY, N., A. GIURGIU, S. JASTRZEBSKI, B. MORRONE, Q. DE LAROUSSILHE, A. GESMUNDO, M. ATTARIYAN, und S. GELLY: *Parameter-efficient transfer learning for NLP*. In K. CHAUDHURI und R. SALAKHUTDINOV (Hrsg.), *Proceedings of the 36th International Conference on Machine Learning*, Bd. 97 d. Reihe *Proceedings of Machine Learning Research*, S. 2790–2799. PMLR, 2019. URL `https://proceedings.mlr.press/v97/houlsby19a.html`.

[8] PFEIFFER, J., A. KAMATH, A. RÜCKLÉ, K. CHO, und I. GUREVYCH: *AdapterFusion: Non-destructive task composition for transfer learning*. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, S. 487–503. Association for Computational Linguistics, Online, 2021. URL `https://aclanthology.org/2021.eacl-main.39`.

[9] NEHRING, J., N. FELDHUS, und A. AHMED: *Adapters for resource-efficient deployment of nlu models*. In C. DRAXLER (Hrsg.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2023*, S. 217–224. TUDpress, Dresden, 2023. URL `https://www.essv.de/pdf/pdf/2023_217_224.pdf`.

ENTWICKLUNG EINES VOICEBOT-FRAMEWORKS AM BEISPIEL VERKEHRSAUSKUNFT FÜR EINEN NAHVERKEHRSBETRIEB

Lia Frischholz[1], *Lisa Winkler*[1], *Christian Gaida*[1], *Melanie Schindler*[2], *Rico Petrick*[2], *Felix Gräßer*[1]

[1]*alphaspeech c/o Linguwerk GmbH,* [2]*LASA - Lausitz Advanced Scientific Applications gGmbH*
lia.frischholz@alphaspeech.de

Zusammenfassung: Intelligente sprach- und textbasierte Dialogsysteme kommen heute bereits vielfach zum Einsatz. Zugrunde liegende Ansätze reichen dabei von einfachen state-basierten Dialogsystemen bis hin zur Nutzung von Large Language Models (LLM). Das vorgestellte modulare Voicebot-Framework integriert neben der Spracherkennungs-Engine *alphaspeech pro* das Open Source Framwork *Rasa* für Dialogmanagement und Natural Language Understanding (NLU). Am Beispiel eines Verkehrsauskunft-Voicebot wird gezeigt, dass auch ohne LLM eine flexible Dialoggestaltung und damit nutzerfreundliches Voicebot-Verhalten möglich sind.

1 Background

Chat- und Voicebots, d.h. text- und sprachbasierte Dialogsysteme, stehen heute bereits in vielen Anwendungen, z.B. zur Informationsbeschaffung, zur Verfügung. Grundlagen für solche Systeme sind weitreichende Entwicklungen im Bereich automatischer Spracherkennung (ASR) und Natural Language Understanding (NLU). Sowohl für ASR [1, 2] als auch im Bereich Intent-Erkennung und Entität-Extraktion [3, 4] – als zentrale Funktionalitäten für NLU – stehen leistungsfähige Technologien und Modelle zur Verfügung. Künstliche neuronale Netze zur Sequenz-zu-Sequenz-Modellierung und dabei insb. Encoder-Decoder- und Transformer-Architekturen sind heute State of the Art [5]. Die Verfügbarkeit großer Sprach-(Large Language Model; LLM) und Foundation Modelle (FM) ermöglichen zusätzlich ganz neue Ansätze und Möglichkeiten für NLU [6]. Darüber hinaus existieren heute mächtige Conversational AI Frameworks (CAIFs), mit denen sich Chat- und Voicebots mit reduziertem Entwicklungsaufwand realisieren lassen [7].

2 Methode und Daten

Im Rahmen eines Proof-of-Concepts soll ein generisches Dialogsystem (Voicebot- Framework) am Beispiel Verbindungsauskunft für einen Nahverkehrsbetrieb konzipiert, entwickelt und erprobt werden. Die Anforderungen an das System sind die Ermöglichung einer (a) multimodalen (text- und sprachbasiert), (b) intuitiven (natürlich-sprachlichen) und (d) effizienten (kurze Dialoge, robuste ASR und NLU) Informationsbeschaffung. Außerdem soll eine einfache Übertragung auf andere Anwendungs-Szenarien möglich sein. Diese Anforderungen resultieren in einer modularen Systemarchitektur, bestehend aus den voneinander unabhängigen und über Standardschnittstellen (APIs) einfach austauschbaren Einzelkomponenten ASR, NLU und TTS.

Als ASR kommt die *alphaspeech pro* Engine [8] mit Adaption des Language Models (LM) an den Anwendungsfall zum Einsatz und für die Ausgabe natürlichsprachlicher Antworten die Open-Source-Sprachsynthese *OpenTTS* [9]. Für die NLU-Funktionalitäten sowie Dialogmanagement wird das CAIF *Rasa* [10] eingesetzt.

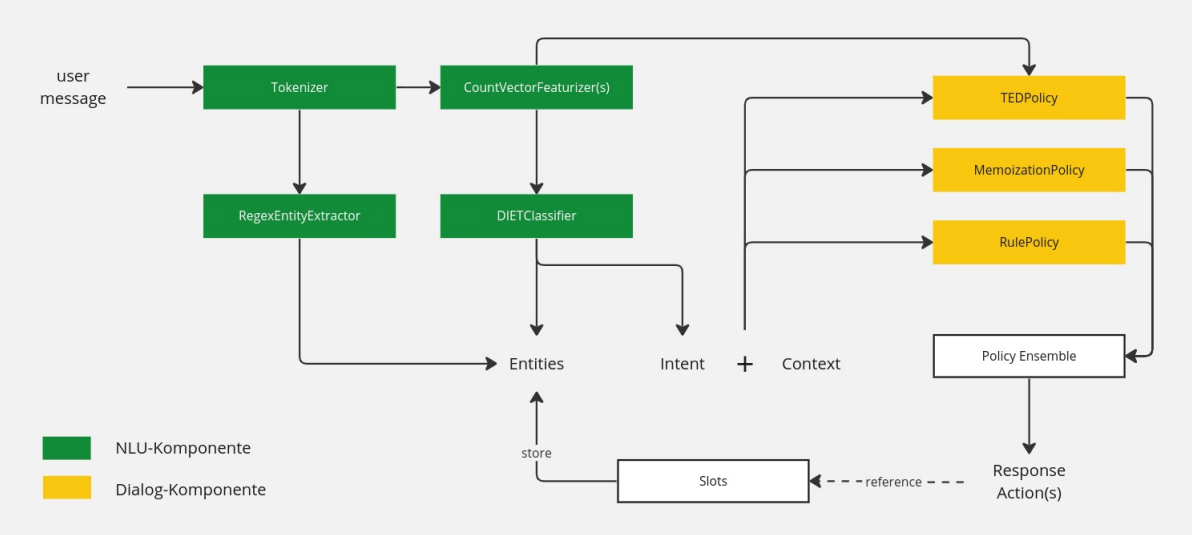

Abbildung 1 – Zusammenspiel der NLU- und Dialog-Komponenten des Voicebot-Frameworks.

Mit dem Rasa-Framework lassen sich NLU und Dialogsteuerung anwendungsspezifisch definieren, konfiguieren und Modelle trainieren. Das Zusammenspiel der genutzten Komponenten ist im folgenden beschrieben und in Abb. 1 veranschaulicht.

Als NLU-Komponenten zur Intent-Erkennung und Entitäten-Extraktion wird hier neben einer regex-basierten Entitäten-Extraktion der State-of-the-Art-Klassifikator „Dual Intent and Entity Transformer“ (DIET) [4] genutzt. Auf Basis eines Korpus aus Beispieläußerungen mit annotierten Intents und Entitäten wird dieser für die Anwendung trainiert. Das Korpus dient gleichzeitig automatisiert zur LM-Adaption.

Das Dialogmanagement wird duch das Zusammenspiel von drei Dialog-Komponenten, sog. „Policies“, realisiert: Für Intents, auf die eine vorgegebene Aktion folgen muss, werden Regeln definiert. Mit dieser „Rule Policy“ lassen sich traditionelle lineare bzw. state-basierte Dialoge abbilden. Die „Memoization Policy“ nutzt ein Korpus aus Beispieldialogen, mit welchen der aktuelle Dialog abgeglichen wird. Kann ein übereinstimmender Dialog gefunden werden, wird die nächste Aktion im Beispieldialog ausgeführt. Greifen weder „Rule Policy“ noch „Memoization Policy“, kommt die Transformer-basierte „TED Policy“ („Transformer Embedding Dialogue“) zum Einsatz, welche anhand von Beispieldialogen die wahrscheinlichste nächste Aktion für einen Kontext lernt [11]. Dieses Zusammenspiel ermöglicht eine nicht-lineare und flexible Dialogführung, die auf Anwendungsdaten beruht und damit als natürlich empfunden wird [11].

Sämtliche Nutzerinteraktionen werden automatisch gespeichert, sodass eine kontinuierliche Optimierung von Dialogen, NLU und LM auf Basis von Anwendungsdaten erfolgen kann.

3 Ergebnisse und Diskussion

Es steht ein Voicebot für das adressierte Anwendungsszenario als Demonstrator zur Verfügung. Besonders von der aufeinander abgestimmten Adaption von LM und NLU- Modellierung profitiert die Robustheit von ASR und NLU und damit Nutzerzufriedenheit und -akzeptanz. Diese Adaption kann weitestgehend automatisiert und kontinuierlich auf Basis von Anwendungsdaten erfolgen. Ein Dashboard zur Verwaltung und Analyse von erfassten Interaktionen und Nutzer-Feedback erleichtert den Optimierungsprozess zusätzlich. Dank der modularen Architektur und dieser Entwicklungs-Werkzeuge lässt sich das Dialogsystem schnell auf weitere Anwendungs-Szenarien übertragen und für diese iterativ optimieren. Die Varianz im Interaktionsverhalten von Testnutzern zeigte, dass (1) frühzeitige Nutzerintegration und (2) Einschluss eines breiten Testnutzerspektrums für den Entwicklungs- und Optimierungsprozess entscheidend sind.

Literatur

[1] BAEVSKI, A., H. ZHOU, A. MOHAMED, and M. AULI: *wav2vec 2.0: A framework for self-supervised learning of speech representations*. 2020. 2006.11477.

[2] RADFORD, A., J. W. KIM, T. XU, G. BROCKMAN, C. MCLEAVEY, and I. SUTSKEVER: *Robust speech recognition via large-scale weak supervision*. 2022. 2212.04356.

[3] CHEN, Q., Z. ZHUO, and W. WANG: *Bert for joint intent classification and slot filling*. 2019. 1902.10909.

[4] MANTHA, M.: *Introducing diet: state-of-the-art architecture that outperforms fine-tuning bert and is 6x faster to train*. 2020. URL https://rasa.com/blog/introducing-dual-intent-and-entity-transformer-diet-state-of-the-art-performance-on-a-lightweight-architecture/.

[5] VASWANI, A., N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, L. KAISER, and I. POLOSUKHIN: *Attention is all you need*. 2023. 1706.03762.

[6] PAASS, G. and S. GIESSELBACH: *Foundation models for natural language processing – pre-trained language models integrating media*. 2023. 2302.08575.

[7] ARONSSON, J., P. LU, D. STRÜBER, and T. BERGER: *A maturity assessment framework for conversational ai development platforms*. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, SAC '21. ACM, 2021. doi:10.1145/3412841.3442046. URL http://dx.doi.org/10.1145/3412841.3442046.

[8] ALPHASPEECH: *alphaspeech pro*. 2024. URL https://alphaspeech.de/pro/. Accessed: 2024-02-18.

[9] OPENTTS: *Open-source text-to-speech*. 2024. URL https://github.com/synesthesiam/opentts. Accessed: 2024-02-18.

[10] RASA: *Open-source conversational ai platform*. 2024. URL https://rasa.com/. Accessed: 2024-02-18.

[11] WHITE, K.: *Unpacking the ted policy in rasa open source*. 2020. URL https://rasa.com/blog/unpacking-the-ted-policy-in-rasa-open-source/. Accessed: 2024-02-17.

MULTIMODALE INTERAKTIVE ASSISTENZ ZU ERHEBUNG VON PATIENT-REPORTED OUTCOME MEASURES

Stefan Hillmann, Philipp Harnisch
Quality and Usability Lab, Technische Universität Berlin
stefan.hillmann@tu-berlin.de

1 Einleitung

Die patientenzentrierte Versorgung gewinnt zunehmend an Bedeutung im Gesundheitswesen, wobei Patient-reported Outcome Measures (PROMs) eine Schlüsselrolle bei der Beurteilung des Behandlungserfolgs aus Patientensicht spielen [1, 2, 3]. PROMs bieten wertvolle Einblicke in die Lebensqualität und das Wohlbefinden der Patienten, stoßen jedoch aufgrund technischer und organisatorischer Herausforderungen oft auf Hindernisse bei der Erhebung als auch der Integration in klinische Abläufe [4]. Das Projekt MIA-PROM adressiert diese Herausforderungen durch die Entwicklung eines multimodalen Assistenzsystems, das den digitalen Sammelprozess vereinfacht und personalisiert. Durch innovative Technologien und ein partizipatives Design strebt das Projekt danach, die Effizienz und Genauigkeit bei der Erfassung von PROMs zu erhöhen und damit die Patientenversorgung nachhaltig zu verbessern.

2 Erhebung von PROMs und die Herausforderungen

PROMs haben sich zu einem unverzichtbaren Werkzeug in der modernen Gesundheitsversorgung entwickelt, da sie direkte Einblicke in die von Patienten wahrgenommene Qualität ihrer Behandlung und Lebensqualität bieten. Diese Maße reflektieren die subjektive Sicht der Patienten auf ihren Gesundheitszustand, ihre Symptome und den Einfluss von Erkrankungen auf ihr tägliches Leben. Der Wert von PROMs liegt vor allem in ihrer Fähigkeit, die Patientenversorgung zu personalisieren und die Behandlungsergebnisse aus der Perspektive der Patienten zu bewerten. Sie ermöglichen es Gesundheitsdienstleistern, die Effekte medizinischer Interventionen besser zu verstehen und Behandlungspläne anzupassen, um den individuellen Bedürfnissen der Patienten gerecht zu werden.

Trotz ihrer offensichtlichen Vorteile stehen die Sammlung und Integration von PROMs in die klinische Praxis vor Herausforderungen [5, 4, 6]. Zu diesen gehören technische Barrieren, wie die Notwendigkeit spezifischer Softwarelösungen und digitaler Plattformen, um Daten effizient zu erfassen und zu analysieren. Organisatorische Hindernisse umfassen die Integration von PROMs in bestehende klinische Abläufe und Informationssysteme sowie die Sicherstellung der Datensicherheit und des Datenschutzes. Ein weiteres zentrales Problem ist die Patientenbeteiligung; es kann schwierig sein, Patienten zur regelmäßigen und ehrlichen Beantwortung von Fragebögen zu motivieren, besonders wenn sie den Eindruck haben, dass ihre Rückmeldungen keinen Einfluss auf ihre Behandlung haben. Zudem erschweren kulturelle und sprachliche Unterschiede die Entwicklung von PROMs, die für alle Patientengruppen gleichermaßen relevant und verständlich sind.

Die Überwindung dieser Herausforderungen erfordert innovative Ansätze, die Technologie, Patientenbetreuung und organisatorische Anpassungen integrieren. Lösungen könnten die Entwicklung benutzerfreundlicher digitaler Werkzeuge für die Erfassung von PROMs sowie (integrierte) Strategien zur Steigerung der Patientenbeteiligung umfassen.

3 Das MIA-PROM System

Das MIA-PROM System steht im Zentrum unseres Projekts zur Verbesserung der Sammlung von PROMs. Dieses innovative System verwendet künstlicher Intelligenz, um eine benutzerfreundliche und effiziente Erfassung von PROMs zu ermöglichen. Es nutzt multimodale Interaktionsmöglichkeiten, insb. Sprache, GUI und Touch-Eingaben, um den unterschiedlichen Bedürfnissen und Präferenzen der Patienten gerecht zu werden. Darüber hinaus bietet das System personalisierte Assistenzdienste, die Patienten durch den Prozess der Dateneingabe leiten und dabei helfen, Missverständnisse und Fehler zu minimieren. Durch den Einsatz von virtuell und physisch verkörperten Avataren, schafft unser System eine interaktive und ansprechende Umgebung, die die Patientenbeteiligung fördert [7, 8, 9, 10].

Durch den Einsatz von KI-gestützten Technologien bietet das System personalisierte Interaktionen, die die Patienten durch den Prozess der Erhebung führen, potenzielle Unklarheiten klären und motivierende Rückmeldungen geben. Ein zentrales Merkmal ist die Fähigkeit, die Interaktion basierend auf dem Verhalten und den Rückmeldungen der Nutzer zu optimieren. Beispielsweise kann das System erkennen, wenn ein Patient Schwierigkeiten mit bestimmten Fragen hat und zusätzliche Informationen oder eine vereinfachte Fragestellung anbieten. Darüber hinaus kann das Userinterface und der Kommunikationsstil angepasst werden, um eine angenehmere und zugänglichere Erfahrung für verschiedene Benutzergruppen zu schaffen.

Die Konzeption und Entwicklung des Systems erfolgt in enger Zusammenarbeit mit Mitgliedern der Zielgruppen – einschließlich Patient:innen, medizinischem Personal und Personen mit Expertise im Bereich Gesundheitstechnologie – entwickelt, um sicherzustellen, dass es den realen Bedürfnissen und Anforderungen entspricht. Diese partizipative Designmethode ermöglichte es, wertvolles Feedback direkt in den Entwicklungsprozess einfließen zu lassen und das System entsprechend anzupassen [11, 12]. Die Einbeziehung eines Beirats von Betroffenen gewährleistete, dass die Perspektiven und Erfahrungen der Nutzenden im Mittelpunkt der Entwicklung stehen. Dadurch soll die Nutzerfreundlichkeit und Akzeptanz des Systems maßgeblich verbessert werden.

4 Schlussfolgerung und Ausblick

Der Ansatz und die Entwicklungen im Projekt MIA-PROM bieten bedeutende Einblicke und Methoden für die Digitalisierung der Gesundheitsversorgung und die Erhebung von PROMs. Durch die Priorisierung von Benutzerfreundlichkeit und Zugänglichkeit zeigt MIA-PROM, wie digitale Werkzeuge gestaltet werden können, um den unterschiedlichen physischen und kognitiven Fähigkeiten von Menschen in der Rehabilitation gerecht zu werden.

Darüber hinaus dient der partizipative Entwicklungsprozess, wie er im Projekt angewendet wird und einen Patientenbeirat einschließt, als Modell für einen inkludierendes und empathisches Entwicklungsprozess im digitalen Gesundheitswesen. Diese Methode stellt sicher, dass die entwickelten Lösungen nicht nur technisch effizient, sondern auch tiefgreifend abgestimmt auf die spezifischen Bedürfnisse und Präferenzen der jeweiligen Zielgruppe sind. Die in MIA-PROM gewonnen Erkenntnisse können zukünftig als Leitfaden dienen, um inklusivere und effektivere digitale Gesundheitslösungen zu schaffen.

5 Fördervermerk

Das Verbundprojekt MIA-PROM wird durch das Bundesministerium für Bildung und Forschung (BMBF) gefördert. Die Arbeiten an der TU Berlin, im Rahmen von MIA-PROM, finden unter dem Förderkennzeichen 16SV9018 statt.

References

[1] DEAN, S., F. AL SAYAH, and J. A. JOHNSON: *Measuring value in healthcare from a patients' perspective*. *Journal of Patient-Reported Outcomes*, 5(S2), pp. 88, s41687–021–00364–4, 2021. doi:10.1186/s41687-021-00364-4.

[2] KAWSKI, S. and U. KOCH: *Zum Stand der Qualitätssicherung in der Rehabilitation Zur Entwicklung der medizinischen Rehabilitation in den 90er-Jahren: Zur Entwicklung der medizinischen Rehabilitation in den 90er-Jahren*. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, 45(3), pp. 260–266, 2002. doi:10.1007/s00103-002-0382-7.

[3] FARIN, E. and W. JÄCKEL: *Qualitätssicherung und Qualitätsmanagement in der medizinischen Rehabilitation*. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, 54(2), pp. 176–184, 2011. doi:10.1007/s00103-010-1206-9.

[4] KÖHN, S., A. SCHLUMBOHM, M. MARQUARDT, A. SCHEEL-SAILER, S. TOBLER, J. VONTOBEL, and L. MENZI: *Predicting non-response in patient-reported outcome measures: results from the Swiss quality assurance programme in cardiac inpatient rehabilitation*. *Int. J. Quality in Health Care*, 34(4), p. mzac093, 2022. doi:10.1093/intqhc/mzac093.

[5] STEINBECK, V., S.-C. ERNST, and C. PROSS: *Patient-Reported Outcome Measures (PROMs): ein internationaler Vergleich: Herausforderungen und Erfolgsstrategien für die Umsetzung von PROMs in Deutschland*. Tech. Rep., Bertelsmann Stiftung, 2021. doi:10.11586/2021053.

[6] BEIERLEIN, V. and H. SCHULZ: *Ergebnismessung in der orthopädischen Rehabilitation*. 2020. URL `https://www.vpksh.de/fileadmin/user_upload/Website/Studien-Gutachten/Q4d_Ergebnismessung_in_der_orthopaedischen_Rehabilitation_Gesamtbericht.pdf`.

[7] BOUMANS, R., F. VAN MEULEN, K. HINDRIKS, M. NEERINCX, and M. OLDE RIKKERT: *A Feasibility Study of a Social Robot Collecting Patient Reported Outcome Measurements from Older Adults*. *International Journal of Social Robotics*, 12(1), pp. 259–266, 2020. doi:10.1007/s12369-019-00561-8.

[8] KRUMMHEUER, A. L., M. REHM, and K. RODIL: *Triadic Human-Robot Interaction. Distributed Agency and Memory in Robot Assisted Interactions*. In *Comp. 2020 ACM/IEEE Int. Conf. on HRI*, pp. 317–319. ACM, Cambridge United Kingdom, 2020. doi:10.1145/3371382.3378269.

[9] SEVERINSON-EKLUNDH, K., A. GREEN, and H. HÜTTENRAUCH: *Social and collaborative aspects of interaction with a service robot*. *Robotics and Autonomous Systems*, 42(3-4), pp. 223–234, 2003. doi:10.1016/S0921-8890(02)00377-9.

[10] HÖFLICH, J. R.: *Relationships to Social Robots: Towards a Triadic Analysis of Media-oriented Behavior*. *intervalla*, 1, pp. 35–48, 2013.

[11] VON UNGER, H.: *Partizipative Gesundheitsforschung: Wer partizipiert woran? Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, Vol 13(1), p. No 1 (2012): Participatory Qualitative Research, 2012. doi:10.17169/FQS-13.1.1781.

[12] CLEMENSEN, J., S. B. LARSEN, M. KYNG, and M. KIRKEVOLD: *Participatory Design in Health Sciences: Using Cooperative Experimental Methods in Developing Health Services and Computer Technology*. *Qualitative Health Research*, 17(1), pp. 122–130, 2007. doi:10.1177/1049732306293664.

VOICE INTERACTION IN MOTION: EAASY VUI AND PHYSICAL EXERTION

Matthias Busch[1], *Long Nguyen*[1], *Ingo Siegert*[1]

[1]*Mobile Dialog Systems, Institute for Information Technology and Communications, Otto von Guericke University Magdeburg, Germany*
(firstname.lastname)@ovgu.de

1 Introduction

Voice User Interfaces (VUIs) present diverse potential for application in bicycle logistics processes [1]. The Eaasy Project, focused on developing a semi-autonomous cargo bike for integration into local courier, express, and parcel (CEP) services, incorporates a Voice User Interface (VUI) designed to facilitate better access and control of the cargo bike.

A critical aspect of this study involves examining how the use of voice technology varies under different levels of physical exertion. Existing research, such as [2], [3], [4], and [5], has explored the influence of mental or physical load on speech signals, particularly in the context of VUIs under physical stress. [6] also investigates the impact of physical exertion on breathing during speech production. However, these studies predominantly focus on read speech samples, with spontaneous, unprompted samples receiving less attention.

This study introduces the "Eaasy VUI Exertion" experiment, investigating the use of the Eaasy VUI under physical strain. The experiment is divided into two phases: The first phase examines the articulation and word choice across various intents, while the second phase simulates the use of a VUI in the context of package delivery. We want to present the design of each experiment phase as well as first insights into the collected dataset.

2 Generation of suitable Intent Images

We explore the use of AI-generated images with symbols to represent intentions, aiming to reduce the time required compared to manually creating visual representations [7, 8]. Our online survey, completed by 39 participants (38% male, 62% female, average age 24.55), assessed the recognizability and reproducibility of these images. While 90% were German speakers, proficiency varied. Results show AI-generated images effectively convey intentions, though limited variability in formulations was noted. Promising images will be used in the follow-up stress study.

3 Physical Exertion in spontaneous speech

The quantification of stress using electrocardiography (ECG) was paired with the recording of verbal commands using intention images, as explored in a preliminary study. Phase 1 focused on spontaneous formulations under stress while walking on a treadmill, encompassing three stress levels (rest, moderate, and maximal) with five recorded verbal commands each. In Phase 2, spontaneity was relinquished in favor of predefined commands to enhance comparability, simulating a more realistic scenario of stress induced by carrying packages while climbing stairs. The study was conducted from November 2023 to January 2024, involving 30 participants with

an average age of 24.2 years, 36.67% female and 63.33% male, all native German speakers. Evaluation encompassed prosody, content, and automatic speech recognition (ASR) errors.

During both the resting phase and maximal exertion, comparisons of speech parameters were conducted, including mean pitch, maximum intensity, and utterance duration. The hypothesis posited that during maximal exertion, speech would be higher in pitch, louder, and shorter in duration. However, absolute values indicated minimal change in utterance length compared to read speech experiments [3], while percentage changes on the graph suggested significant deviations, with alterations in utterance length measured in milliseconds. Mean pitch and maximum intensity exhibited considerable variation among participants, with some even speaking quieter during maximal exertion. Changes in volume were observable, with a shift of 10 dB due to the logarithmic nature of sound perception, though this varied among individuals.

Acknowledgment

The content and results of this work were developed within the framework of the research project Electric Adaptive Autonomous Smart deliverY System (Eaasy System, Reference: 01ME21004E), funded by the Federal Ministry for Economic Affairs and Climate Action.

Literaturverzeichnis

[1] BUSCH, M., M. KANIA, T. ASSMANN, and I. SIEGERT: *Radlogistik als Anwendungsgebiet für digitale Sprachassistenten – Ein Diskussionsbeitrag*. In *Elektronische Sprachsignalverarbeitung 2023. Tagungsband der 34. Konferenz*, vol. 105 of *Studientexte zur Sprachkommunikation*, pp. 223–230. TUDpress, München, Germany, 2023.

[2] SCHULLER, B., F. FRIEDMANN, and F. EYBEN: *The munich biovoice corpus: effects of physical exercising, heart rate, and skin conductance on human speech production*. In *Proceedings 9th Language Resources and Evaluation Conference, LREC '14*, Studientexte zur Sprachkommunikation, pp. 1506–1510. Reykjavik, Iceland, 2014.

[3] GODIN, K. W. and J. H. L. HANSEN: *Analysis of the effects of physical task stress on the speech signal*. *Journal of the Acoustical Society of America*, pp. 3992–3398, 2011.

[4] USMAN, M.: *On the performance degradation of speaker recognition system due to variation in speech characteristics caused by physiological changes*. In *International Journal of Computing and Digital Systems*, vol. 6, pp. 119–127. Saudi Arabia, 2017.

[5] ENTWISTLE, M. S.: *The performance of automated speech recognition systems under adverse conditions of human exertion*. *International Journal of Human–Computer Interaction*, 16, pp. 127–140, 2013.

[6] TROUVAIN, J. and R. WERNER: *Muster der Sprechatmung in verschiedenen Sprechstilen – Eine Pilotstudie*. In C. DRAXLER (ed.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2023*, pp. 95–102. TUDpress, Dresden, 2023. URL `https://www.essv.de/pdf/pdf/2023_95_102.pdf`.

[7] GAUVIN, H. S., M. K. JONEN, J. CHOI, K. MCMAHON, and G. I. DE ZUBICARAY: *No lexical competition without priming: Evidence from the picture–word interference paradigm*. *Quarterly Journal of Experimental Psychology*, 71(12), pp. 2562–2570, 2018.

[8] CARR, T. H., C. MCCAULEY, R. D. SPERBER, and C. M. PARMELEE: *Words, pictures, and priming: on semantic activation, conscious identification, and the automaticity of information processing. Journal of Experimental Psychology: Human perception and performance*, 8(6), p. 757, 1982.

First Steps into ASPIRE: A Pilot Study on Automated Speech Analysis Regarding Psychotherapeutic Alliance in Psychotherapies

Martha Schubert[1,2,4], Michael Schenk[2,4], Julia Krüger[2,4], Melanie Elgner[2,4], Florian Junne[2,3,4], Ingo Siegert[1]

[1] *Mobile Dialog Systems Group, Faculty of Electrical Engineering and Information Technology, Otto von Guericke University, Magdeburg, Germany,* [2]*Department of Psychosomatic Medicine and Psychotherapy, Medical Faculty, Otto von Guericke University, Magdeburg, Germany,* [3]*German Center for Mental Health (DZPG), partner site Halle-Jena-Magdeburg, Jena, Germany,* [4] *Center for Intervention and Research on adaptive and maladaptive brain - Circuits underlying mental health (C-I-R-C), Halle-Jena-Magdeburg*
martha.schubert@ovgu.de

Abstract: In the ASPIRE pilot study, audio and video recordings are gathered from psychotherapy sessions at the Department of Psychosomatic Medicine and Psychotherapy Magdeburg. Aiming to correlate speech markers with the therapeutic alliance, as a robust predictor of treatment success, acoustic prosodic as well as linguistic markers are extracted from both patient and therapist speech.

1 Background

The quality of the relationship between therapists and patients in psychotherapies, known as the therapeutic alliance, is one of the most important predictors of treatment success [1]. In the course of technical progress in automated speech analysis, studies reveal correlations between the therapeutic alliance and the way patients and therapists talk to each other. In our pilot study, ASPIRE, we aim to deepen and expand these findings by analysing correlations between ratings of the therapeutic alliance and specific speech markers as well as the similarity of patient's and therapist's speech.

2 Methods

In our study, we evaluate acoustic prosodic, as well as linguistic, speech markers. Acoustic prosodic markers include, e.g. frequency and intensity features of the voices, rhythmic measures such as speaking and articulation rate, speech quality markers such as shimmer, jitter, and the HNR (harmonic to noise ratio). To extract linguistic markers, the recordings are transcribed using state-of-the-art offline automatic speech recognition engines. We then apply the LIWC (Linguistic Inquiry and Word Count) software, which calculates measures like Language Style Matching [2]. Finally, we combine linguistic and acoustic prosodic speech markers to calculate the similarity of the patient's and the therapist's speech.

Currently, the data is being collected. We record psychotherapy sessions in audio and video at the Department of Psychosomatic Medicine and Psychotherapy in Magdeburg. The therapeutic alliance is rated by patients and therapists after each session using the Working Alliance Inventory (WAI) [3]. Furthermore, it is planned to add objective ratings by independent raters using a rating system [4] for detecting alliance ruptures and reparations during therapy sessions. Up to now, around 60 therapy sessions with a duration of 25 to 50 minutes each were recorded.

3 Exemplary Preliminary Results

While the process of collecting the data is still ongoing, some preliminary analyses from one fully recorded psychotherapy of a therapeutic dyade can already be shown:

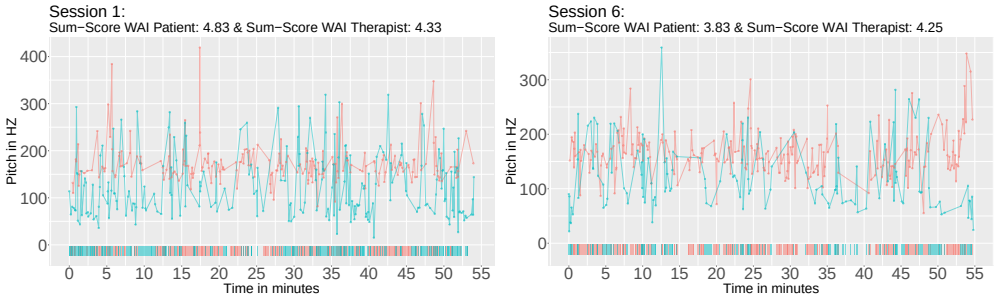

Figure 1 – Session 1 (left) and Session 6 (right), showing pitch (top) and turn-taking (bottom) of therapist and patient

As shown in figure 1, session 6 exhibits extended pauses and higher pitch variation during dialogue compared to session 1. Furthermore, the patient rated the therapeutic alliance better in session 1 (3.83 vs. 4.83). Further analysis is needed to test whether this indicates a statistically relevant positive correlation.

4 Conclusion

Connecting speech markers with therapeutic alliance (as a first psychtherapeutic model construct) reveals promising initial findings. Our research is led by the future vision of a real-time feedback system for psychotherapists, that helps to predict psychotherapeutic process and outcome by using speech analysis [5].

References

[1] FLÜCKIGER, C., A. C. DEL RE, B. WAMPOLD, and A. HORVATH: *The alliance in adult psychotherapy: A meta-analytic synthesis*. *Psychotherapy*, 55, 2018. doi:10.1037/pst0000172.

[2] AAFJES-VAN DOORN, K., J. PORCERELLI, and L. MÜLLER-FROMMEYER: *Language style matching in psychotherapy: An implicit aspect of alliance*. *Journal of Counseling Psychology*, 67, pp. 509–522, 2020. doi:10.1037/cou0000433.

[3] WILMERS, F., T. MUNDER, R. LEONHART, T. D. P. HERZOG, R. PLASSMANN, J. BARTH, and H. W. LINSTER: *Die deutschsprachige version des working alliance inventory – short revised (wai-sr) – ein schulenübergreifendes, ökonomisches und empirisch validiertes instrument zur erfassung der therapeutischen allianz.* 2008. URL `https://api.semanticscholar.org/CorpusID:147642307`.

[4] EUBANKS, C. and J. MURAN: *Rupture resolution rating system (3rs): Manual version 2022*. 2023. doi:10.13140/RG.2.2.29780.17282.

[5] KRÜGER, J., I. SIEGERT, and F. JUNNE: *Künstliche Intelligenz für die Sprachanalyse in der Psychotherapie – Chancen und Risiken*. *PPmP - Psychotherapie · Psychosomatik · Medizinische Psychologie*, 72(09/10), p. 395–396, 2022. doi:10.1055/a-1915-2589. URL `http://dx.doi.org/10.1055/a-1915-2589`.

KONVERSATIONELLE INTERAKTION FÜR HYBRIDE VERANSTALTUNGEN

Stefan Schaffer & Aaron Ruß
DFKI Labor Berlin
stefan.schaffer@dfki.de

ZUSAMMENFASSUNG: Hybride Veranstaltungen zeichnet aus, dass sie sowohl vor Ort zum Beispiel in Kongress- oder Veranstaltungsräumen als auch online auf virtuellen Plattformen stattfinden. Seit der Corona Pandemie haben zahlreiche Anbieter virtueller Veranstaltungsplattformen ihre Angebote weiterentenwickelt und um Funktionalitäten erweitert[1]. Durch den virtuellen Anteil sind bei hybriden Veranstaltungen immer mehr digitale Daten verfügbar. Neben Informationen über die Agenda, den Veranstaltungsort, oder häufig gestellter Fragen sind zunehmend während des Events auch Echtzeitdaten sowie Informationen aus Social Media verfügbar. Im Projekt ToHyVe (**To**olbox für **hy**bride **Ve**ranstaltungsformate) werden digitale Informationen über hybride Veranstaltungen über einen auf Vorarbeiten basierenden Backend-Service für konversationelle Interaktion [1] bereitgestellt und über Nutzerschnittstellen von Veranstaltungsplattformen integriert. In unserem Beitrag berichten wir über die nutzerzentrierte Entwicklung des sprachbasierten konversationellen Dienstes, die auf Retrieval Augmented Generation (RAG) [2] basierende Architektur des Systems und über Vorgehensweisen und Erkenntnisse zum Prompt Engineering mit Vicuna [3] und anderen LLM. Abschließend geben wir einen Ausblick über die weiteren geplanten Arbeiten unserer Forschung, welche untere anderem die Evaluation des Gesamtsystems beinhalten.

NUTZERZENTRIERTE ENTWICKLUNG: Konversationelle Interaktion soll über einen so genannten Concierge Chatbot als Dienst realisiert werden, so dass unterschiedliche Anwendungen im Kontext hybrider Veranstaltungen einfach sprach- oder textbasierte Eingabeformate integrieren können. Für die nutzerzentrierte Gestaltung des konversationellen Concierge wurde ein Ideation-Workshop durchgeführt. Aus den Workshop-Ergebnissen wurden die möglichen Themengebiete des Concierge herausgearbeitet. Es sollen vier Module entwickelt werden für „Agenda“, FAQs“, „Exponate“ und für „Vortragsinhalte“. Die Module werden im nächsten Abschnitt näher erläutert.

ARCHITEKTUR: Der konversationelle Dienst funktioniert indem bei einem ersten Schritt eine Klassifizierung der Nutzeräußerung vorgenommen wird. Je nach Klassifizierung wird dann eines von drei Modulen im Dienst für die tatsächliche Beantwortung verwendet, entweder auf Grundlage von (1) strukturierten Daten einer Agenda (also Fragen zum Zeitplan, Ort oder Vortragenden einer Veranstaltung) oder (2) halb-strukturierte Daten zu Frequently Asked Questions (FAQ, d.h. häufige Fragen/Antworten zu einer Veranstaltung) oder (3) basierend auf unstrukturierten Daten zu Exponaten oder Vortragsinhalten. Um eine größere Wissensbasis für Exponatsbeschreibungen und Vortragsinhalten für die Beantwortung von Nutzeräußerungen verfügbar zu machen, wird eine Retrieval Augmented Generation (RAG) -Architektur verwendet, bei welcher Textausschnitte passend zu der Nutzeräußerung herausgesucht werden [2]. Diese werden dann in die Anweisungen für das Large Language Model (LLM) eingebettet, wofür das generative Sprachmodell Vicuna [3] (basierend auf LLaMA2 [4]) zum Erzeugen der Antwort verwendet wird. Die Komponenten der Sprachverarbeitungskette werden mittels dem Framework LangChain[2] eingebunden, wobei der RAG-Ansatz in Form des Hybride List Aware Transformer Reranking (HLTR [5]) realisiert wurde. Für die Umsetzung wird im ersten Schritt die Vector Store-Implementierung von Facebook AI Similarity Search (FAISS [6]) verwendet, um anhand von Embedding-Vektoren eine Vorauswahl an passenden Textstücken zu suchen. Diese werden dann in einem nächsten Schritt mit einem BERT-basierten Cross Encoder [7] verarbeitet, um dann in einem letzten Schritt mit einem Reranking eine kleine Anzahl der bestpassenden Textausschnitte zu finden, die dann in den Anweisungs-Prompt für das LLM eingebettet werden. Auf diese Weise werden größere Mengen von unstrukturiertem Text als Wissensbasis für die Beantwortung von Nutzerfragen verfügbar gemacht. Das Auffinden und Einbetten von passenden Textausschnitten, die relevanten Inhalte zur Beantwortung der Nutzeräußerung enthalten, hilft dabei, die berüchtigten Halluzinationen von generativen Sprachmodellen zu reduzieren.

[1] Z.B. www.tricat.net oder www.iventic.com
[2] https://github.com/langchain-ai/langchain

ERKENNTNISSE: Die bisherigen Ergebnisse sind ausschließlich qualitativ. Die Ausgabe des Systems soll in deutscher Sprache sein. Testungen mit Prompts in englischer und deutscher Sprache haben ergeben, dass die Antworten des Concierge besser sind, wenn Instruktionen im Prompt auf Englisch formuliert werden. Für nicht beantwortbare Fragen wurde das Modell eingangs auf Englisch instruiert zu antworten, dass es keine Antwort weiß. Als Resultat wurden Antworten wie „Ich don't know" generiert. Das Problem konnte behoben werden indem in die englischsprachige Instruktion ein expliziter Antwortsatz auf Deutsch integriert wurde. Die allgemeine Struktur der Promts in der Form „Instruktion – Kontext (aus RAG) – Few shot examples" funktioniert gut. Vergleichstests mit anderen open-source Modellen haben ergeben, dass diese Modelle mehr halluzinieren als das verwendete Modell, selbst wenn der Kontext mitgegeben wird. Testungen mit Promt Injections ergaben, dass das Modell durchaus anfällig ist.

AUSBLICK: Wir planen aktuell eine Prozedur zum automatischen Testen der Auswirkungen von Veränderungen der Promts. Des weiteren wollen wir Methoden zur Reduzierung der Möglichkeit von Promt Injections integireren.

AUTOREN

Stefan Schaffer ist Senior Researcher und Leiter der Gruppe Mensch-KI-Interaktion am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI), Abteilung Kognitive Assistenzsysteme

Aaron Ruß ist Senior Software Engineer und leitender Entwickler der Gruppe Mensch-KI-Interaktion am Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI), Abteilung Kognitive Assistenzsysteme

REFERENZEN

[1] Schaffer, S., Ruß, A., Sasse, M. L., Schubotz, L., & Gustke, O. (2021). Questions and answers: important steps to let AI chatbots answer questions in the museum. In International Conference on ArtsIT, Interactivity and Game Creation (pp. 346-358).

[2] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.

[3] **Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality** by: The Vicuna Team, besucht am 12.12.2023-

[4] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

[5] Zhang, Y., Long, D., Xu, G., & Xie, P. (2022). HLATR: enhance multi-stage text retrieval with hybrid list aware transformer reranking. arXiv preprint arXiv:2205.10569.

[6] Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P. E., ... & Jégou, H. (2024). The faiss library. arXiv preprint arXiv:2401.08281.

[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

POTENTIALS OF CHATBOTS IN THE GERMAN PUBLIC ADMINISTRATION

Oliver Jokisch, Karl M. Walter
Digital Administration Dept., HSF Meissen University of Public Administration
oliver.jokisch@hsf.sachsen.de

MOTIVATION: A chatbot (originally chatterbot [1]) simulates human conversation via text or voice interactions. Simpler applications exist for decades, and recent variants, based on (generative) artificial intelligence (AI), enable conversations close to natural language, cf. ChatGPT [2]. Chatbot interfaces represent an easy way to perform various tasks like information search or smart-home control, and some of them are well established in everyday live. Throughout the world, the business is discovering the potential of underlying speech and AI technologies as a customer-oriented instrument in numerous applications. In parallel, the adopters of such technologies need to consider legal hurdles, such as e.g. the sovereignty of personal data, anchored in the European General Data Protection Regulation (EU 2016/679, GDPR) [3]. Typical public administrations, however, are conservative in deploying new technologies, although they are urged to improve their services and internal processes in terms of efficiency and user friendliness. Related to German administrations, a law to improve the online access to the services from 2017 (Online Access Act [4]) obligated states and municipalities to offer all relevant services electronically (originally by December 2022), supplemented by an online access, to make the services more attractive and efficient for citizens and companies. The original roadmap of the Online Access Act failed [5], but with respect to use cases including chatbot interfaces, some good practices in administration could be demonstrated, such as the chatbot Kora since 2018 [6].

METHOD: The study deals with the questions, whether voice or text-based chatbots can effectively support legal and ergonomic requirements (e.g. correctness, barrier-free access, quicker response), and which major challenges in the administrative deployment occur. The approach bases on empirical findings by asking for preferable application scenarios of voice interfaces in administration, and by comparing the strengths, weaknesses, opportunities and threats (SWOT method). The concept includes an online survey covering different aspects of a voice-assistant usage, the potentials of using them in administration, and the experience participants already have with such solutions [7], embracing both, pre-defined multiple-choice questions as well as free-text replies in German. For completion, we discuss technical means to deal with the raised issues and illustrate experiments with a text-based chatbot for administrational use cases [8], such as filling a form to apply for parental allowance, or an FAQ bot, supplemented by a SWOT analysis and the search of already existing chatbots in the public sector.

RESULTS: The survey [9] was distributed among employees in different Saxon administrations, at which 81 people participated, by answering 15 questions. The replies indicate a strong interest in voice assistants, but also a sceptic mindset, related to the technology and solutions (as known to the participants) and to a potential application in administrative environments. The results show potential obstacles for voice interfaces in the administrative deployment, in which 45.7 % of the nominations refer to a missing trust in the technology or security. The user friendliness and the acceptance among the staff are doubted by 42 % of the participants. In contrast, only 18.5 % of the answers are concerned about useful tasks for a voice assistant, or the costs. The free-text replies confirm the quantitative results, and some accent the importance of privacy. The advantage of the (voice-based) chatbots is primarily seen at internal, repetitive tasks rather than for the external communication with citizens, which should stay a human task. Despite some successful (text-based) chatbot implementations in federal states and municipalities, such solutions are still receiving less consideration than, e.g. in the federal government, universities or insurance companies since ca. 10 years [8]. The related potential analysis confirms the internal application scenarios, e.g. HR, IT service desk, or communication between authorities.

CONCLUSIONS: The study shows the potential of voice and text-based chatbots in administration. The embedding of chatbots is mostly seen in holistic approaches of agile work or digitalization. In the external communication with citizens or companies, form or FAQ bots open interesting functional and user-oriented potential. With respect to the SWOT analysis, a significant increase in efficiency and

reduction of complexity can be achieved. The risks and weaknesses can be reduced through professional process management (as usual in software introduction), precise process modeling, a detailed tender, a consultation of data protection officers, and by a sufficient training of employees.

Oliver Jokisch is professor of cybersecurity at the Meissen University of Public Administration (HSF) in Germany and the director of the Saxon Institute for Governance Innovation (SIVIM). Oliver studied information technology at the TU Dresden and the Loughborough University (UK), graduating as a diploma engineer. He holds a Ph.D. degree from TU Dresden and previously had a chair of system theory at the private university of Deutsche Telekom in Leipzig. His current research focuses on the digitalization in the public sector and on AI applications in audio, speech, and video processing.

Karl M. Walter is an alumnus of the Digital Administration Department at the Meissen University of Public Administration (HSF). Karl wrote his thesis under the supervision of Prof. Oliver Jokisch and Prof. Gunnar Auth (same department), and he is working at a legal authority in Saxony.

LIST OF REFERENCES

[1] MAULDIN, M.: ChatterBots, TinyMuds, and the Turing Test: Entering the Loebner Prize Competition. Proc. 12th National Conference on Artificial Intelligence, Seattle, pp. 16-21, August 1994. Retrieved https://cdn.aaai.org/AAAI/1994/AAAI94-003.pdf , 18/02/2024.

[2] OPENAI: Introducing ChatGPT (Optimizing language models for dialogue, version 3.5), 30 November 2022. Retrieved from https://openai.com/blog/chatgpt/ , 18/02/2024.

[3] REGULATION (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, 27 April 2016. Retrieved from https://gdpr.eu/tag/gdpr/ , 31/01/2024.

[4] FEDERAL MINISTRY OF THE INTERIOR, BUILDING AND COMMUNITY, GERMANY: "Gesetz zur Verbesserung des Onlinezugangs zu Verwaltungsleistungen" (Onlinezugangsgesetz, Online Access Act), August 2017, retrieved https://www.gesetze-im-internet.de/ozg/ , 31/01/2024.

[5] INITIATIVE D21 E. V. / KANTAR GMBH: eGovernment Monitor 2022, October 2022, retrieved https://initiatived21.de/app/uploads/2022/10/egovernment_monitor_2022.pdf , 18/02/2024.

[6] CITY OF HEIDENHEIM: "Integrierter Chatbot – was ist das?" (Kora 3.0), January 2021, retrieved from https://www.heidenheim.de/neue+webseiten+2020#anker912700 , 18/02/2024.

[7] JOKISCH, O.; BRAUNER, K.; SIEGERT, I.: A preliminary study on voice-assisted interfaces in the German public administration. Proc. KM2022 Conf., Ljubljana, p. 42, June 2022.

[8] WALTER, K.: Potential analysis and experiments with a chatbot in administration (in German), Bachelor thesis, Meissen University of Public Administration (HSF), September 2023.

[9] BRAUNER, K.: Case scenarios and risks of a voice assistant in the administrative application (in German), Bachelor thesis, Meissen University of Public Administration (HSF), April 2022.

Index

Busch, Matthias, 22

Dunkelberg, Matthias, 9

Elgner, Melanie, 24

Fingscheidt, Tim, 9
Frischholz, Lia, 16

Gaida, Christian, 16
Gräßer, Felix, 16

Harnisch, Philipp, 19
Hillmann, Stefan, 19

Jokisch, Oliver, 28
Junne, Florian, 24

Klein, Andreas M., 6
Krüger, Julia, 24

Leschanowsky, Anna, 12
Li, Zhengyang, 9
Lohrenz, Timo, 9

Merkel, Sebastian, 4

Nehring, Jan, 14
Nguyen, Long, 22

Peters, Nils, 12
Petrick, Rico, 16
Popp, Birgit, 12

Ruß, Aaron, 26

Schaffer, Stefan, 26
Schenk, Michael, 24
Schindler, Melanie, 16
Schubert, Martha, 24

Siegert, Ingo, 22, 24

Walter, Karl M., 28
Werner, Steffen, 7
Wienrich, Carolin, 5
Winkler, Lisa, 16