Anouck Braggaar

Tilburg University
Warandelaan 2
5037 AB Tilburg

A.R.Y.Braggaar@tilburguniversity.edu
https://www.tilburguniversity.edu/staff/a-r-y-braggaar

1 Research interests

Many companies use **dialogue systems** for their **customer service**, and although there has been a rise in the usage of these systems (Costello and LoDolce, 2022), many of these systems still face challenges in comprehending and properly responding to the customer (Følstad et al., 2021). In our project[1] we aim to figure out how to develop and improve these conversational agents. Part of this project, focuses on the detection of **breakdown patterns** and the possible solutions (**repairs**) to mitigate negative results of these errors.

1.1 Conversational breakdowns

Breakdowns lead to frustration and an overall downgraded customer experience (Ashktorab et al., 2019). Therefore it is important to be able to detect these breakdowns and properly solve them to mitigate these negative effects. One of the important questions to start with when looking at breakdowns is to define what a breakdown actually is and what triggers this breakdown (or taking a different perspective, what happens within conversations without breakdowns?) In the next section I will first discuss the research plan we have to figure out if there are different kinds of breakdowns and eventually in Section 1.2 if we can solve the consequent issues through repairs. In this project we will focus on text-based task-oriented customer service chatbots; incorporating features such as speech will lead to very different breakdowns (for example arising from the ASR part of a system).

Errors are often the cause of leading to a breakdown, which leads to the user not being able to continue the conversation (Higashinaka et al., 2015b). There have been attempts to create taxonomies of errors for open-domain systems (Higashinaka et al., 2015a, 2021). Similar to our project, the work of Reinkemeier and Gnewuch (2022) focuses on a text based dialogue system in a specific domain (in their case an insurance company). They aim to find the causes of conversational breakdowns by conducting a cluster analysis of messages leading to breakdowns. We will follow a similar approach as Reinkemeier and Gnewuch (2022) by trying to cluster utterances and figure out if we can detect reasons for initiating repairs. We use real-life Dutch chatbot data from a railroad company. The conversations cover a diverse range of topics, from asking for a ticket refund to travel directions.

Are there any linguistic patterns to be found in utterances before breakdowns occur? Or are there certain topics the chatbot is not capable of handling? To figure out the potential reasons for breakdowns, we use repairs as a proxy. The advantage of using the railroad chatbot dataset is that it has a fixed set of chatbot initiated repairs. From this set we have selected three general repairs that are used in various situations:

1. Not understanding the user and asking for rephrasing: 'Unfortunately I don't fully understand what you mean. Could you rephrase the question in different words? Tip: I understand short and concise questions the best.'
2. Not being able to help and redirecting to human employee: 'I'm sorry, I believe I can not help you yet. Shall I connect you with my colleague?'
3. Apologising and redirecting to human employee: 'I'm sorry to hear that something isn't to your satisfaction. I can unfortunately not register your complaint, but my colleague from customer support is happy to help. Click on the button below.'

These repairs are used anytime the chatbot is not capable of answering the customer query (the last focuses on complaints but is also used in situations were the customers is slightly negative). Possibly not all breakdowns/miscommunications are caught with this approach (for example when the chatbot answers with an irrelevant answer) but the dataset is too large to manually examine every conversation.

Similar to Reinkemeier and Gnewuch (2022) we will use a clustering approach to figure out if there are patterns to be found in breakdowns. We will add multiple features partly derived from Reinkemeier and Gnewuch (2022) who use for example semantic weight and percentage of unknown words. For example, we will also use the number of sentences, characters, and tokens in an utterance. We also will create more complex features as well. As an example we will make use of commonness as described by Meij et al. (2012). Making use of anchors, this metric scores commonness of n-grams based

[1] https://www.conversationalagentsresearch.com/

Proceedings of the 19th Annual Meeting of the Young Reseachers' Roundtable on Spoken Dialogue Systems, pages 26–29
September 11–12, 2023.

on Wikipedia data. We will combine this score together with training data of the bot. This means that words with high scores for commonness, that are not part of the training data, might indicate a wrong interpretation.

1.2 ... and Repairs

Miscommunication is an important concept in human language (see for an extensive discussion for example Healey et al. (2018)), sometimes resulting in breakdowns. It is not always possible to prevent breakdowns, which underscores the importance of repairs. As breakdowns occur in many different situations it is necessary to critically think about the 'best' repair for any given situation. So, after focusing on breakdowns we like to find out how to mitigate these breakdowns by using repairs. As was discussed in Section 1.1 we have used the existing repairs as a proxy to detect breakdowns. We could wonder if these repairs are actually the best repairs to fix a conversational breakdown. Different forms of breakdowns, systems or different user groups might need different repair strategies. Ashktorab et al. (2019) for example discuss that chit-chat systems have different goals with repairs (not repairing but engaging for further conversation).

Repair is an important notion studied in conversation analysis to study problem resolving in conversation. The basis of the notion is explained by Schegloff et al. (1977). The notion of repair is later also applied on dialogue systems as breakdowns in conversation with bots are common. Ashktorab et al. (2019) investigate user preference for eight repair strategies. Some of these strategies occur in commercial systems, others are novel strategies that incorporate some of the inner workings of the algorithms behind the dialogue system. They find that both providing options and giving explanations are preferred by users (Ashktorab et al., 2019). Bohus and Rudnicky (2005) focus on non-understanding errors and recovery strategies in spoken systems. They compare the recovery strategies and also investigate how the user responds to these strategies. A different approach is taken by Cuadra et al. (2021) who investigate the self repair of a spoken system (Amazon Alexa) and how it affects the interaction. They show that if an error occurs, a repair is appreciated but when no error occurs a repair can worsen the experience. Lastly, Skantze (2005) examine how humans recover from speech recogniser errors by corrupting speech output. These errors will be similar in spoken dialogue systems. They show that if participants face speech recognition errors, they will ask task-related questions.

2 Spoken dialogue system (SDS) research

Since my submission last year, much has changed within the field of dialogue systems. With the advent of ChatGPT and subsequent open alternatives (such as Alpaca (Taori et al., 2023) and Open Assistant), there has been renewed (media)attention for dialogue systems and chatbots. These new technologies will bring new possibilities for research into dialogue systems but also new challenges. I suspect that much more research will focus on the challenges and problems these systems will bring, in for example the context of education (Kasneci et al., 2023) and hospitality (Gursoy et al., 2023). I also suspect that the (general) public gets more and more familiar with these systems and the (assumed) capabilities of systems like chatGPT. Due to both positive and negative attention to these technologies, expectations of the public towards dialogue systems will also shift. Therefore it seems important to learn more about expectations of users and the ways in which we can manage those expectations. Previous research has already shown that expectation management is an important factor. For a chatbot to be successful the user needs to know what to expect from the beginning (Brandtzaeg and Følstad, 2018). Previous work has also stressed the importance of understanding the user perceptions and expectations before building the chatbot (Zamora, 2017), and creating chatbots with characteristics that are in line with users' expectations (Chaves and Gerosa, 2021). Users tend to evaluate chatbots worse when the experience does not line up with their expectations (de Sá Siqueira et al., 2023). Similar research is now also done with chatGPT, for example surveying the expectations of healthcare workers on adopting chatGPT in their work (Temsah et al., 2023).

3 Suggested topics for discussion

Breakdowns and repairs Can we mitigate negative effects after encountering erroneous chatbots with only repairs or are there other solutions as well? Should we tailor repairs to specific situations or breakdowns?

Cooperate with industry In what way can academia cooperate with industry and how far should we go to make our research usable directly for industry? For which purposes can our research be used in the industry? Research has already shown that big tech companies shape research to cater to their needs (Whittaker, 2021; Abdalla and Abdalla, 2021), having its influence grow over the last few years (Abdalla et al., 2023).

Incorporating ChatGPT What are the issues with incorporating current technologies like ChatGPT in dialogue systems for both research and industry? Can we overcome issues with interpretability, transparency and replicability? How should we evaluate closed models if we don't know what is exactly in the training data (Rogers et al., 2023)? Should our focus be on the more open models such as Stanford Alpaca (Taori et al., 2023)?

Acknowledgements

This project has been funded by the NWO Smooth Operators project (KIVI.2019.009).

References

Mohamed Abdalla and Moustafa Abdalla. 2021. The grey hoodie project: Big tobacco, big tech, and the threat on academic integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. pages 287–297.

Mohamed Abdalla, Jan Philip Wahle, Terry Ruas, Aurélie Névéol, Fanny Ducel, Saif M Mohammad, and Karën Fort. 2023. The elephant in the room: Analyzing the presence of big tech in natural language processing research. In *61st Annual Meeting of the Association for Computational Linguistics*.

Zahra Ashktorab, Mohit Jain, Q Vera Liao, and Justin D Weisz. 2019. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. pages 1–12.

Dan Bohus and Alexander I. Rudnicky. 2005. Sorry and I didn't catch that! - an investigation of non-understanding errors and recovery strategies. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*. Special Interest Group on Discourse and Dialogue (SIGdial), Lisbon, Portugal, pages 128–143. https://aclanthology.org/2005.sigdial-1.14.

Petter Bae Brandtzaeg and Asbjørn Følstad. 2018. Chatbots: changing user needs and motivations. *interactions* 25(5):38–43.

Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction* 37(8):729–758.

Katie Costello and Matt LoDolce. 2022. Gartner predicts chatbots will become a primary customer service channel within five years. [Accessed June 14, 2023]. https://www.gartner.com/en/newsroom/press-releases/2022-07-27-gartner-predicts-chatbots-will-become-a-primary-customer-service-channel-within-five-years.

Andrea Cuadra, Shuran Li, Hansol Lee, Jason Cho, and Wendy Ju. 2021. My bad! Repairing intelligent voice assistant errors improves interaction. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW1):1–24.

Marianna A de Sá Siqueira, Barbara CN Müller, and Tibor Bosse. 2023. When do we accept mistakes from chatbots? The impact of human-like communication on user experience in chatbots that make mistakes. *International Journal of Human–Computer Interaction* pages 1–11.

Asbjørn Følstad, Theo Araujo, Effie Lai-Chong Law, Petter Bae Brandtzaeg, Symeon Papadopoulos, Lea Reis, Marcos Baez, Guy Laban, Patrick McAllister, Carolin Ischen, et al. 2021. Future directions for chatbot research: an interdisciplinary research agenda. *Computing* 103(12):2915–2942.

Dogan Gursoy, Yu Li, and Hakjun Song. 2023. Chatgpt and the hospitality and tourism industry: an overview of current trends and future research directions. *Journal of Hospitality Marketing & Management* pages 1–14.

Patrick G. T. Healey, Jan P. de Ruiter, and Gregory J. Mills. 2018. Editors' introduction: Miscommunication. *Topics in Cognitive Science* 10(2):264–278. https://doi.org/https://doi.org/10.1111/tops.12340.

Ryuichiro Higashinaka, Masahiro Araki, Hiroshi Tsukahara, and Masahiro Mizukami. 2021. Integrated taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. pages 89–98.

Ryuichiro Higashinaka, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, Yuka Kobayashi, and Masahiro Mizukami. 2015a. Towards taxonomy of errors in chat-oriented dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Prague, Czech Republic, pages 87–95. https://doi.org/10.18653/v1/W15-4611.

Ryuichiro Higashinaka, Masahiro Mizukami, Kotaro Funakoshi, Masahiro Araki, Hiroshi Tsukahara, and Yuka Kobayashi. 2015b. Fatal or not? Finding errors that lead to dialogue breakdowns in chat-oriented dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 2243–2248. https://doi.org/10.18653/v1/D15-1268.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103:102274.

Edgar Meij, Wouter Weerkamp, and Maarten De Rijke. 2012. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining*. pages 563–572.

Fabian Reinkemeier and Ulrich Gnewuch. 2022. Designing effective conversational repair strategies for chat-

bots. In *Proceedings of the 30th European Conference on Information Systems (ECIS 2022)*.

Anna Rogers, Niranjan Balasubramanian, Leon Derczynski, Jesse Dodge, Alexander Koller, Sasha Luccioni, Maarten Sap, Roy Schwartz, Noah A. Smith, and Emma Strubell. 2023. Closed ai models make bad baselines. https://hackingsemantics.xyz/2023/closed-baselines/.

Emanuel A Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language* 53(2):361–382.

Gabriel Skantze. 2005. Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication* 45(3):325–341.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html* 3(6):7.

Mohamad-Hani Temsah, Fadi Aljamaan, Khalid H Malki, Khalid Alhasan, Ibraheem Altamimi, Razan Aljarbou, Faisal Bazuhair, Abdulmajeed Alsubaihin, Naif Abdulmajeed, Fatimah S Alshahrani, et al. 2023. Chatgpt and the future of digital health: A study on healthcare workers' perceptions and expectations. In *Healthcare*. MDPI, volume 11, page 1812.

Meredith Whittaker. 2021. The steep cost of capture. *Interactions* 28(6):50–55.

Jennifer Zamora. 2017. I'm sorry, dave, i'm afraid i can't do that: Chatbot perception and expectations. In *Proceedings of the 5th international conference on human agent interaction*. pages 253–260.

Biographical sketch

Anouck Braggaar is a second year PhD candidate at Tilburg University. Her work focuses on conversational agents for customer service and is part of the Smooth Operators project[1]. Currently she is working on a literature review on evaluation approaches for task-oriented dialogue systems and on a study to automatically detect reasons for repair. Previously, she received a research master in Linguistics at the University of Groningen.