Voicing Concerns: User-Specific Pitfalls of Favoring Voice over Text in Conversational Recommender Systems

Alain D. Starke[1,2], Minha Lee[3]

[1]*Wageningen University & Research, Droevendaalsesteeg 4, 6708 PB Wageningen, The Netherlands*

[2]*University of Bergen, P.O. Box 7800, 5020 Bergen, Norway*

[3]*Eindhoven University of Technology, Groene Loper 3, 5612 AE Eindhoven, The Netherlands*

Abstract

In the context of Conversational Recommender Systems (CRSs) and Conversational User Interfaces (CUIs; e.g., Digital Assistants, such as Siri), an increasing number of voice-based applications are emerging, often at the expense of text-based applications. In this position paper, we argue that the possible first-mover advantage of adopting voice-based technologies may put specific groups of users at a profound disadvantage, as they are likely to run into accessibility issues. For example, users that stammer or whom are not fluent in the English language have a hard time using voice-based conversational recommender systems. Along this line, we describe a number of challenges and issues for current and future systems.

Keywords

Conversational User Interfaces, Recommender Systems, Accessibility, Inclusion, Voice-based Systems

1. Introduction

Voice-based technologies are diffusing through society at a fast pace. Reportedly 4.2 billion digital voice assistants were in use in 2020 [1], including well-known technologies such as Amazon Alexa and Siri. Their role in the 'Internet of Things' system is becoming increasingly important [2], in the sense incumbent technologies, such as recommender systems, are often made compatible with voice-based applications [3].

One aspect of voice-based or conversational user interfaces is to retrieve personalized content. To date, however, most conversational recommender systems (CRSs) to date are text-based [4]. They focus on mining textual user input, such as through fixed messages in clickable menus or by open-ended text queries [5, 6]. In comparison, the number of voice-based conversational recommender systems are still limited, but is likely to expand in the coming years [3].

The user-system dynamic between text-based and voice-based interactions differs greatly. Whereas text-based CRSs can rely on either open-ended queries or fixed input (e.g., the user selects an answer option), voice-based queries tend to be impromptu and are more complex to process. Nonetheless, given the current share and expected growth of digital assistant use [1], the emergence of digital assistants, such as Amazon Alexa and Google Home, suggests that designing for voice-based interac-

tions will be a bigger target for many commercial applications than text-based systems. For one, a specific application of voice-based interactions for recommender systems research is the users' ability to retrieve personalized suggestions by voice, as hands-free interaction [3, 7].

Despite the possibilities of voice-based interactions, we see some challenges. Specifically, the trend of the commercial landscape that prioritizes voice first comes with disadvantages for specific users who are either not equipped to work with the technology (e.g., Siri) or who are not the targeted, 'mainstream' user.

We briefly give an overview of text and voice systems before we jump to the critique of voice-based recommender systems. We bring up why text-based solutions may be more beneficial in certain contexts and for specific users, making it important text-based CRSs are not discontinued. However, due to the growing trend of voice-based interactions, e.g., the rise of Alexa, we believe that we cannot avoid designing for different types of voice-based interactions in the coming future. For the latter, we will formulate a few suggestions.

2. Conversational Systems

2.1. Text-based systems

Text-based conversational systems, which are also known as chatbots, have been around for decades, such as Weizenbaum's ELIZA in the 1960s [8]. Chatbots now exist on many business-to-consumer websites, for example as an automated customer service agent [9]. In terms of technical implementation, two approaches are taken to build chatbots, which typically also applies to

text-based conversational recommenders [6]. They are either built as command-based systems that respond to user queries as if they are commands or as bots that use natural language understanding. For example, on Slack, one can issue commands (e.g., *unsubscribe*) that chatbots can easily understand rather than using natural language (e.g., "please get me out of this channel") that can be vague for systems to understand [10]. Also, people may easily misspell or give incomplete input that chatbots cannot accurately interpret.

An important challenge of conversational systems is to mitigate misunderstandings or a conversational breakdown [11]. There are inadequate responses to user requests, false positives [11], either because of unclear query or a missing database category [12, 13]. In such cases, conversational repair strategies becomes important like how the system should correct for misunderstood phrases or unclear user intentions [12]. An example of a repair strategy is a system giving potential options that people can choose from, such as "I did not understand that. Did you mean X or Y?", to keep the conversation going.

In sum, the two strategies then are to design 1) command-based systems that allow for minimum flexibility on user input for efficiency or 2) natural language-based systems that allow for greater input flexibility, but also then, with an increased number of possible repair strategies that are not always successful.

2.2. Voice-based systems

Voice-based conversational user interfaces (VUIs) are becoming more popular in everyday use. In particular, smart home assistants such as Google Home and Amazon Alexa are used more frequently to help its users find content that they are looking whether [14], regardless of whether the query is mundane and factual (e.g., 'What weather is it today?'), or more exploratory (e.g., 'Play me a song for my dinner party'). The latter is more commonly explored in recommender systems research, for it seeks to retrieve an item that a user does not explicitly know about.

Current voice-based systems face a number of technical issues that are often situational. For example, considering the use of voice-based assistant in a car [7], there may be environmental noise, people not formulating clearly due to multi-tasking driving and voice interaction, among other causes. Technical issues that come with VUIs are many, and will also impact the design of voice-based conversational recommender systems. To list three, they at times lack noise robustness, multimodal understanding, and addressee detection [7]. In most contexts, users' environmental conditions will feature a certain degree of background noise, such as when people are away from home. Moreover, voice-based system cannot understand multimodal cues like gestures or gaze that accompany users' speech [15], which is likely to affect the interpretability of the voice-based query. Finally, when there is more than one person talking, the system has to distinguish whose voice to zone in on [7], which might lead to conflicts of agency [16]. Each of these problems are challenging. Yet, even if these technical issues are resolved, they will not create a headway for a seamless experience for individuals; inclusion is not about one-size-fits-all, but about how these technical issues do not disproportionally affect specific userse.

3. Critique of Voice-based systems

Voice-based queries tend to be 'messier' than text-based inputs [6]. The current state-of-the-art in Natural Language Processing methods opens up possibilities for more open-ended conversational strategies in recommendation. However, even NLP-based systems are often still limited to familiar input, i.e., requiring an explicit understanding of users' messages, which often gets misunderstood. Problems like of environmental noise distortion of user input are common [7]. Yet when it comes to user adoption, voice-based technologies have diffused at a large speed in terms of innovation adoption [17, 2]. These have both positive and negative side effects.

On the one hand, it seems that voice-based applications be integrated in a modular way, as they can work with recommendation libraries without designing an appropriate user interface. On the other hands, it seems that innovation in technology may only benefit those that can work it. Even for people who are considered to be "regular users", there is a lot of trial and error when it comes to learning how to interact with voice-based agents [18] and recommender systems. But there are people who have additional difficulties due to various differences in abilities; technical solutions often are built around the normative assumption that users are fully able-bodied, i.e., with sight, hearing, and other abilities intact [19].

In recommender domains that use more traditional interfaces, marginalized people are often 'served' by a simple fix. For example, a tourism recommender system for people with physical disabilities would apply post-filtering to an appropriate set of recommendations [20]. The problem for conversational recommenders, however, not only applies to the appropriateness of the suggested context, but also to the usability of the technology in the first place. For example, people who stammer, being a small subset of the population, face difficulties at the start of their interaction: a voice-based system often cannot understand what they say due to the lack of training data and design choices [21]. Their speech does not fit the normative template of how people should "normally"

talk. Furthermore, many smart home assistants are only compatible with a few languages (e.g., they are 'biased' towards English [14]), and speech recognition may be distorted because of fluctuations in human emotions [4, 22]. We realize that these issues on inclusion in the use of voice-based assistants is nuanced [16]. This means that notions on who can easily use voice-based agents depends on multiple factors, such as accents, speech patterns, and the access to commercial agents like Alexa, which introduces many ways that efforts to include can also be exclusive, e.g., by prioritizing one accent over others.

4. Suggestions for Conversational Recommender Systems

We have highlighted different challenges for both text-based and voice-based interactions. What stands out is that some challenges are easier to resolve with user training or adaptation (e.g., lacking sufficient technical knowledge to use a text-based interface), than other challenges (e.g., non-native users lack vocal skils, such as because of stammering) [21]. What these challenges have in common is how people's assumptions about conversational agents, be they chatbots or Alexas, shape their interactions. People may have expectations that conversational agents cannot meet, as the systems cannot yet to complex tasks such as email management by voice [23]. Even text-based chatbots often do not meet people's needs, as users expect a higher level of understanding from bots that they were not designed for [10]. Hence, for most of us, going beyond simple interactions and towards more complex exchanges is a problem that we all share due to the state of the technology.

Some studies describe that conversational recommender systems are distinct from the more traditional chatbots and dialogue-based systems [6]. However, we argue that the retrieval of conversational elements in conjunction with 'task-related items' are two sides of the same coin. A task-based conversation can be dialogue-based, by supporting a task at hand. Instead of focusing on a false dichotomy between task-based or dialogue-based systems, a better way forward is being attentive to how different users' capacities get highlighted or ignored by systems. The problem to focus on is inclusion vs. exclusion of user groups based on systems' assumptions of different abilities that people may or may not have.

How should we move forward with conversational recommender systems? Recommender systems are traditionally applied in domains where one-shot recommendations are effective [24], such as movies, e-commerce, and books. The use of conversations, however, makes for more complex interactions which introduces greater technical challenges. We above differentiated between

text-based chatbots with voice-based agents. In terms of users being "better understood", the decades old text-based interactions may be better suited. Perhaps counter-intuitively, due to the limits of query and text-based conversational recommendation, the odds are smaller that it 'gets it wrong'. Or, more technically, that it generates a negative adversarial response [25], or has a conversational breakdown [12]. Although the usability is arguably lower in the sense that one needs to "touch" an interface, this poses a huge advantage to those who have to concentrate to interact with such a voice-based application. However, the consumer trend is shaping up to favor voice-based applications; IBM, Google, and Amazon have product lines that promote voice-first interactions.

We offer two suggestions moving forward. To optimize accessibility for all users, a move towards 'voice-enabled' rather than 'voice-first' or voice-based recommender systems would be desirable, akin to technology in which 'voice' is a feature rather a key characteristic (e.g., Siri on an Apple iPhone). Although this requires the deployment of two different retrieval and recommendation pipelines, it maximizes accessibility by combining 'the best of two worlds'. To note, we did not consider multimodality, e.g., combination of voice, gaze, body movements, and more, which will become more important in the coming years [26].

We also suggest that diversity of data for retrieval and recommendation is essential to design inclusive conversational recommender systems, or systems that cater to specific users. Efforts are ongoing when it comes to making voice-based interactions more accessible; Google's Project Euphonia[1] aims to collect more data on atypical speech, e.g., from people with cerebral palsy. Similarly, more time should be spend on collecting difficult data when it comes to voice in research, in terms of responding to "unconventional voices".

5. Conclusion

This paper has reflected on current practices in conversational recommender systems. In particular, we have pitted text-based systems against voice-based systems, observing that while voice-based recommender systems are becoming more common because of their integration with digital assistant [3], it may put specific users at a disadvantage. We have identified a number of challenges to make CRSs more inclusive, particularly for the emerging domain of voice-based user interfaces. We emphasize lastly that inclusion for some may mean exclusion for others. In order to recommend to all users, we need to understand all users. Specifically, understanding users not only in terms of preferences, but also in terms of the

---

[1]https://sites.research.google/euphonia/about/

fundamental conversational elements, such as speech, should be a priority.

References

[1] L. S. Vailshery, Number of digital voice assistants in use worldwide from 2019 to 2024 (in billions), 2021. URL: https://www.statista.com/statistics/973815/worldwide-digital-voice-assistant-in-use/.

[2] D. Pal, C. Arpnikanondt, S. Funilkul, W. Chutimaskul, The adoption analysis of voice-based smart iot products, IEEE Internet of Things Journal 7 (2020) 10852–10867.

[3] A. Iovine, F. Narducci, G. Semeraro, Conversational recommender systems and natural language:: A study through the converse framework, Decision Support Systems 131 (2020) 113250.

[4] C. Gao, W. Lei, X. He, M. de Rijke, T.-S. Chua, Advances and challenges in conversational recommender systems: A survey, arXiv preprint arXiv:2101.09459 (2021).

[5] D. C. Hernandez-Bocanegra, J. Ziegler, Conversational review-based explanations for recommender systems: Exploring users' query behavior, in: CUI 2021-3rd Conference on Conversational User Interfaces, 2021, pp. 1–11.

[6] D. Jannach, A. Manzoor, W. Cai, L. Chen, A survey on conversational recommender systems, ACM Computing Surveys (CSUR) 54 (2021) 1–36.

[7] F. Weng, P. Angkititrakul, E. E. Shriberg, L. Heck, S. Peters, J. H. Hansen, Conversational in-vehicle dialog systems: The past, present, and future, IEEE Signal Processing Magazine 33 (2016) 49–60.

[8] J. Weizenbaum, Eliza—a computer program for the study of natural language communication between man and machine, Communications of the ACM 9 (1966) 36–45.

[9] R. Dale, The return of the chatbots, Natural Language Engineering 22 (2016) 811–817.

[10] M. Lee, L. Frank, W. IJsselsteijn, Brokerbot: A cryptocurrency chatbot in the social-technical gap of trust, Computer Supported Cooperative Work (CSCW) 30 (2021) 79–117.

[11] A. Følstad, C. Taylor, Conversational repair in chatbots for customer service: the effect of expressing uncertainty and suggesting alternatives, in: International Workshop on Chatbot Research and Design, Springer, 2019, pp. 201–214.

[12] Z. Ashktorab, M. Jain, Q. V. Liao, J. D. Weisz, Resilient chatbots: Repair strategy preferences for conversational breakdowns, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–12.

[13] K. Corti, A. Gillespie, Co-constructing intersubjectivity with artificial conversational agents: people are more likely to initiate repairs of misunderstandings with agents represented as human, Computers in Human Behavior 58 (2016) 431–442.

[14] B. R. Cowan, P. Doyle, J. Edwards, D. Garaialde, A. Hayes-Brady, H. P. Branigan, J. Cabral, L. Clark, What's in an accent? the impact of accented synthetic speech on lexical choice in human-machine dialogue, in: Proceedings of the 1st International Conference on Conversational User Interfaces, 2019, pp. 1–8.

[15] D. Heylen, Head gestures, gaze and the principles of conversational structure, International Journal of Humanoid Robotics 3 (2006) 241–267.

[16] M. Lee, R. Noortman, C. Zaga, A. Starke, G. Huisman, K. Andersen, Conversational futures: Emancipating conversational interactions for futures worth wanting, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–13.

[17] G. McLean, K. Osei-Frimpong, Hey alexa… examine the variables influencing the use of artificial intelligent in-home voice assistants, Computers in Human Behavior 99 (2019) 28–37.

[18] C. M. Myers, L. F. Laris Pardo, A. Acosta-Ruiz, A. Canossa, J. Zhu, "try, try, try again:" sequence analysis of user interaction data with a voice user interface, in: CUI 2021-3rd Conference on Conversational User Interfaces, 2021, pp. 1–8.

[19] S. Costanza-Chock, Design justice: Towards an intersectional feminist framework for design theory and practice, Proceedings of the Design Research Society (2018).

[20] R. Mahmoud, N. El-Bendary, H. M. Mokhtar, A. E. Hassanien, Similarity measures based recommender system for rehabilitation of people with disabilities, in: The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Beni Suef, Egypt, Springer, 2016, pp. 523–533.

[21] L. Clark, B. R. Cowan, A. Roper, S. Lindsay, O. Sheers, Speech diversity and speech interfaces: Considering an inclusive future through stammering, in: Proceedings of the 2nd Conference on Conversational User Interfaces, 2020, pp. 1–3.

[22] J. Pittermann, A. Pittermann, W. Minker, Emotion recognition and adaptation in spoken dialogue systems, International Journal of Speech Technology 13 (2010) 49–60.

[23] E. Luger, A. Sellen, "like having a really bad pa" the gulf between user expectation and experience of conversational agents, in: Proceedings of the 2016 CHI conference on human factors in computing systems, 2016, pp. 5286–5297.

[24] G. Adomavicius, A. Tuzhilin, Toward the next

generation of recommender systems: A survey of the state-of-the-art and possible extensions, IEEE transactions on knowledge and data engineering 17 (2005) 734–749.

[25] G. Penha, C. Hauff, What does bert know about books, movies and music? probing bert for conversational recommendation, in: Fourteenth ACM Conference on Recommender Systems, 2020, pp. 388–397.

[26] Y. Deldjoo, J. R. Trippas, H. Zamani, Towards multimodal conversational information seeking, in: Proceedings of the ACM Conference on Research and Development in Information Retrieval, SIGIR, volume 21, 2021.