Exploring Users' Preferences for Chatbot's Guidance Type and Timing

Meng-Hsin Wu[*]
menghsin.wu@mail.utoronto.ca
University of Toronto
Toronto, Ontario, Canada

Su Fang Yeh[*]
sfy.iem07g@nctu.edu.tw
National Yang Ming Chiao Tung University
Hsinchu, Taiwan

XiJing Chang
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
siliconcrystal.c@nycu.edu.tw

Yung-Ju Chang[†]
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
armuro@cs.nctu.edu.tw

ABSTRACT

While task-oriented chatbots have become popular recently, conversational breakdowns are still common and will often lead to unfavorable user experiences. Guidance serves a crucial role in helping users to understand how to have better interaction with chatbots. Nonetheless, questions like what kinds of guidance to provide and when to provide guidance remain underexplored. In this study, we examined users' preferences for two types of guidance (Example-Based and Rule-Based) at four guidance timings (Service-Onboarding, Task-Intro, After-Failure, and Upon-Request). Our results show that users preferred Example-based guidance, and generally preferred guidance provided at Task-Intro. Example-based guidance appearing at Task-Intro was the favorite guidance combination for most participants. Through analysis of participants' explanations of their preferences, the strengths and weaknesses of these guidance types and guidance timings are presented. The preliminary results are based on a subset of the data (n=24). Further in-depth investigation into the underlying factors that influence users' preferences for guidance, as well as the interplay effect between guidance type and guidance timing is needed.

ACM Reference Format:
Meng-Hsin Wu, Su Fang Yeh, XiJing Chang, and Yung-Ju Chang. 2021. Exploring Users' Preferences for Chatbot's Guidance Type and Timing. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '21 Companion), October 23–27, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3462204.3481756

1 INTRODUCTION

In recent years, chatbots have drawn attention from various domains. As big-tech companies have provided open source APIs for building bots, tremendous growth in chatbot applications can be seen on many text messaging platforms. There are approximately 300,000 active chatbots on Facebook Messenger [4], spanning multiple domains: e-commerce [1], causal chats[1], travel [16], and many more. While the majority of the chatbots have a commercial focus, past findings have also shown that chatbots can provide opportunities to create positive social impact in many domains [9, 31] such as health care [28], education, [7], supporting systems [3], etc. Nevertheless, despite much research attention on designing chatbots for social applications, a fundamental challenge that users face when interacting with chatbots, reducing or overcoming conversation breakdowns, has not drawn enough attention. Conversation breakdowns or obstacles can be caused by the chatbot expressing misunderstanding of users' messages or not recognizing users' intents and providing unexpected messages [21]. These breakdowns are likely to trigger users' negative emotions [30, 34], and in some severe situations, cause the users to abandon the chatbot service [14, 21].

Research has suggested that people's negative reactions to or the trigger of conversational breakdowns may come from their overestimation of a chatbot's ability to understand their messages [14, 18, 21]. To solve this issue, Weisz et al. [33] proposed a role-play tutorial to increase participants' empathy and to set their realistic expectations for chatbots. When conversation breakdowns happen, the repair strategies that chatbots should adopt have also been investigated [2].

On the other hand, informing users of the chatbot's capability upfront or providing appropriate guidance regarding how to interact with the chatbot may even prevent conversation breakdowns [13, 14, 26]. Nevertheless, despite the continual call for the need to guide users on how to use chatbots, there has been little attention paid to what kind of guidance should be offered. For example, example-based guidance is frequently adopted in various kinds of intelligent user interfaces, perhaps due to its ability to express complex concepts[5, 14, 17, 32]. However, it is suggested that examples make it more difficult for people to infer the underlying rules [8, 29]. Explicitly stating rules, or rule-based guidance, is considered useful when unexpected situations happen [23]. Nonetheless, some studies have revealed that people need to take more time to learn rules than to learn examples[8, 24]. In the context of interacting

[*] Both authors contributed equally to this research.

[1] https://www.facebook.com/chatbots.io

CSCW '21 Companion, October 23–27, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8479-7/21/10…$15.00
https://doi.org/10.1145/3462204.3481756

with chatbots, to the best of our knowledge, little attention has been paid to how guidance type, for instance, the example-based and rule-based approaches, influence users' present performance, as well as their subsequent conversation with the chatbot. It is also unclear how users perceive these types of guidance in helping them interact with the chatbot in terms of user experience.

Finally, the timing of guidance can play a crucial role too, since unexpected guidance often distracts users' attention from their main task and slows them down [27]. However, suggestions of timing have been diverse. For example, Nielsen's heuristic evaluation [25] recommends that guidance should always be accessible for users to get help and documentation; recently, specialized heuristics for conversation agents [19] also suggest that chatbots should guide users by clarifying system capabilities. Jain et al.[14] found that participants preferred guidance in the initial stage of the interaction. Kirschthaler et al. [17] adopted interviews and found that, compared to automatically provided guidance, participants preferred guidance that shows up when requested via using a voice user interface [10]. Nevertheless, thus far, there has not been a formal investigation that compares different timings for offering specific kinds of guidance.

In this paper, we report the results of a preliminary study that investigated users' desirable and undesirable combinations of guidance type and timing. Specifically, we studied two guidance types: Example-Based and Rule-Based, and compared users' preferences for them at four different timings: Service-Onboarding, Task-Intro, After-Failure, and Upon-Request. The key research questions are: 1) When is the ideal timing to provide guidance, 2) Which types of guidance should be provided, and 3) What is the most preferred combination of guidance type and timing?

2 METHOD

2.1 Chatbots and Study Tasks

We developed our chatbot using IBM Watson[2]. Participants came to the lab and interacted with our designed chatbot, performing 12 close-ended interaction tasks with predetermined goals. Interaction tasks were designed based on two dimensions: context and task complexity. We designed two task contexts, one related to arranging travels and the other related to movie booking, both of which have been extensively used in prior chatbot research [2, 14, 15, 20, 22]. We adopted Campbell's notion of task complexity [6], where the more requirements the task involves, the higher the complexity of the task. We divided all tasks into three complexity levels, namely 4, 6, and 8, based on the number of required pieces of information to accomplish the tasks.

2.2 Study Design and Procedure

We employed a two-phase study, consisting of the first-phase between-subjects experiment to obtain quantitative results and the second-phase reflection session to obtain qualitative feedback on different combinations of guidance type and timing. There were nine conditions for phase one, which included 2 (Guidance Types: Example-Based vs. Rule-Based) x 4 (Guidance Timing: Service-Onboarding vs. Task-Intro vs. After-Failure vs. Upon-Request) conditions, and one control condition in which participants received no guidance. Service-Onboarding refers to the timing in which the user receives guidance before entering a specific task. Task-Intro refers to the timing in which the user receives guidance right after entering a specific task. After-Failure timing refers to the timing in which the user receives guidance when a conversational breakdown occurs. Upon-Request timing is when the user actively requests guidance by clicking a help button. The detailed guidance dialogue under each timing can be found in the auxiliary materials.

In each condition, participants performed 12 tasks that were equally broken down into two trials. In the first trial, participants performed six tasks in a partial counterbalancing order [12]. During the conversation, the chatbot offered an assigned guidance combination to assist participants in accomplishing tasks. Participants took a five-minute break before moving on to the second trial, in which they performed the same kinds of tasks, with small variations in the task description and requested information. The purpose was to examine whether participants could successfully accomplish similar tasks based on what they had learned in the first trial. Participants' performance in both trials, including effectiveness (task success rate), efficiency (task execution time), and learnability (efficiency improvement in two trials) were measured. After finishing two trials, participants filled out an online questionnaire about their satisfaction with the guidance and their overall experience of using the chatbot upon completing the tasks.

Participants then moved on to the second-phase reflection study after finishing the first phase. The purpose of the reflection study was to let participants reflect on their perceptions, attitudes, and concerns about each guidance combination, and to analyze their preferences. Participants were shown the eight other guidance features that they were not exposed to in the first phase. We asked them about their preferences and had them rank all nine guidance combinations, and then we used this as a prompt to probe the explanations behind their guidance rankings. We carefully presented each combination in order to assist participants in comparing multiple guidance combinations more easily.

We first showed the same type of guidance they received, but at the other three different timings, of which the order was randomly determined. After comparing the same guidance type under all timings, we had participants compare another guidance type under different timings in a similar fashion. Whenever viewing a new guidance combination, participants were asked to reflect their thoughts and preferences for that combination and modify their ranking if necessary.

2.3 Participants

We recruited 126 participants who were older than 20 years old from multiple social networking platforms, including Ptt, Dcard, and Facebook groups. All participants had provided their backgrounds, including their demographic information, familiarity with chatbots, and familiarity with technology [2], in our sign-up online questionnaire when they were registering for our study. We balanced participants' backgrounds and assigned them into conditions by adopting a semi-randomization approach. That is, while randomly assigning each of the selected participants to one of the nine conditions, we also sought to balance the number of participants and

[2] https://www.ibm.com/watson

their backgrounds across the nine conditions[12]. Each participant received 300 TWD (approximately 10 USD) as compensation for their participation. In this paper, we reported preliminary results on the first 24 participants. These 24 participants were aged between 20 and 39 (M=26). The female-male sex ratio was 1:1.18. While 62.5% of participants had interaction experience with chatbots prior to the study, the rest had no such experience.

3 PRELIMINARY RESULT

3.1 Quantitative Results

Our results show that participants had clear preferences for specific guidance combinations. We computed the preference score of each guidance combination from the 24 participants, with the top one receiving a score of 9 and the last one receiving a score of 1. Overall, the top three combinations receiving the highest score were: Task-Intro Example-Based (M=6.63), Upon-Request Example-Based (M=6.25), and Task-Intro Rule-Based (M=5.67), and the three which received the lowest scores were: no guidance (M=2.54), After-Failure Rule-Based (M=4.00), and After-Failure Example-Based (M=4.25). The average scores of all guidance combinations are shown in Table 1. Generally speaking, Example-Based guidance received higher scores than their counterpart rule-based guidance (i.e. at the same timing), except for the Service-Onboarding timing. Participants generally preferred receiving guidance at Task-Intro the most and After-Failure the least. We conducted the Friedman test [11] on the ranking outcome of nine guidance combinations, and the results indicated that the overall differences in the ranking outcomes were statistically significant (χ^2=39.967, p<.001).

We also analyzed the frequency of each guidance combination being ranked at a specific position using three categories (1st-3rd, 4th-6th, and 7th-9th). This result indicates how often each guidance combination was perceived as being in the top three, middle, or bottom three categories they preferred to see in chatbot interaction. We also observed a similar result: Example-Based guidance was generally preferred by the participants; three out of the four Example-Based type of guidance were ranked in the top three favorite types of guidance were ranked in the top three favorite type of guidance more than 40% of the time; among them, Example-Based guidance presented in the Task-Intro timing even received 54.17% of the vote. In contrast, 50% of the time, participants put guidance After-Failure as their least favorite, regardless of the guidance type; it was only better than providing no guidance at all (control,62.5%).

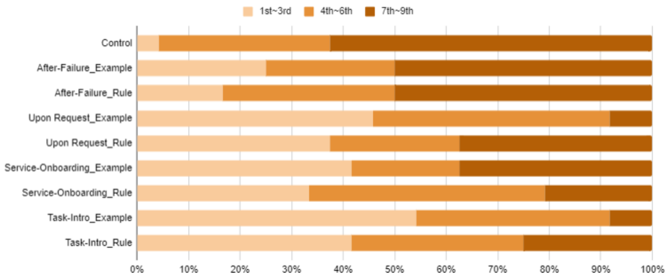

Figure 1: Relative frequency for User Preference Ratings, i.e. percentage for the three rank group, by scenarios.

Table 1: Descriptive Statistics of User Preference Ratings Test

Scenario	Mean	SD
Control	2.54	2.226
Service-Onboarding Example	5.17	2.648
Service-Onboarding Rule	5.42	2.358
Task-Intro Example	6.63	2.203
Task-Intro Rule	5.67	2.426
Upon-Request Example	6.25	1.962
Upon-Request Rule	5.13	2.643
After-Failure Example	4.25	2.364
After-Failure Rule	4.00	2.187

3.2 Qualitative Results

We used affinity diagramming to analyze participants' interview transcripts. The findings emerging from the analysis also tended to explain the quantitative results. Specifically, the perceived two major advantages of Example-Based guidance were its clear scope and easiness to understand and follow. Participants thought the language structure of Example-Based guidance helped them better understand what they could type. Also, some participants reported that Example-Based guidance was easier to follow because they could directly apply it to their own messages, as P19 described, "*if I were provided with example-based guidance, I can just copy and paste, changing 12/10 to 10/10, which won't cause me any burden because I don't have to think. It's more like editing a template.*" With that being said, there were still some participants who preferred Rule-Based guidance more as it served as explicit bullet points, and made participants feel like they could input information in a more flexible way.

Among all timings of Example-Based guidance, Task-Intro was the most favored timing mainly because earlier guidance could reduce the number of failed utterances. Many participants generally disliked how they had to fail the conversation first before they saw guidance. This made them feel that their efforts during interaction were in vain. Service-Onboarding guidance and Task-Intro guidance are both early forms of guidance. However, Service-Onboarding guidance was generally ranked lower because some participants preferred to interact with the chatbot directly without reading the Service-Onboarding introduction.

4 CONCLUSION AND FUTURE WORK

Our preliminary results demonstrate that task-oriented chatbot users have clear preferences for different combinations of guidance type and timing. Specifically, the most popular combination was Task-Intro Example-Based guidance based on the subset data. However, the current work did not examine the interaction between guidance type and guidance timing, i.e., whether any combination of the timings and types would result in better users' performance and satisfaction. Future work can benefit from evaluating these combinations. Our work will consider the entire dataset, including participants' quantitative performance-wise data (i.e., efficiency, effectiveness, and learnability) measured in the first-phase interaction tasks, participants' satisfaction scores, and their preference scores for all guidance combinations.

ACKNOWLEDGMENTS

This research was supported in part by the Ministry of Science and Technology, R.O.C (109-2218-E-009-016).

REFERENCES

[1] Rachel Arthur. 2016. *Burberry Is Also Experimenting With Chatbots For London Fashion Week.* Retrieved June 5, 2021 from https://www.forbes.com/sites/rachelarthur/2016/09/17/burberry-is-also-experimenting-with-chatbots-for-london-fashion-week/?sh=386a969dffd4

[2] 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.* 1–12.

[3] Petter Bae Bae Brandtzæg, Marita Skjuve, Kim Kristoffer Kristoffer Dysthe, and Asbjørn Følstad. 2021. When the Social Becomes Non-Human: Young People's Perception of Social Support in Chatbots. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–13.

[4] Marion Boiteux. 2018. *Messenger at F8 2018.* Retrieved January 16, 2021 from https://blog.messengerdevelopers.com/messenger-at-f8-2018-44010dc9d2ea

[5] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces.* 258–262.

[6] Donald J Campbell. 1988. Task complexity: A review and analysis. *Academy of management review* 13, 1 (1988), 40–52.

[7] Fabio Clarizia, Francesco Colace, Marco Lombardi, Francesco Pascale, and Domenico Santaniello. 2018. Chatbot: An education support system for student. In *International Symposium on Cyberspace Safety and Security.* Springer, 291–302.

[8] Nick Ellis. 1993. Rules and instances in foreign language learning: Interactions of explicit and implicit knowledge. *European Journal of Cognitive Psychology* 5, 3 (1993), 289–318.

[9] Asbjørn Følstad, Petter Bae Brandtzæg, Tom Feltwell, Effie LC Law, Manfred Tscheligi, and Ewa A Luger. 2018. SIG: chatbots for social good. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems.* 1–4.

[10] Asbjørn Følstad and Ragnhild Halvorsrud. 2020. Communicating Service Offers in a Conversational User Interface: An Exploratory Study of User Preferences in Chatbot Interaction. In *32nd Australian Conference on Human-Computer Interaction.* 671–676.

[11] Milton Friedman. 1940. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics* 11, 1 (1940), 86–92.

[12] Frederick J Gravetter and Lori-Ann B Forzano. 2018. *Research methods for the behavioral sciences.* Cengage Learning.

[13] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[14] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N Patel. 2018. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference.* 895–906.

[15] F Kaptein, J Broekens, K Hindriks, and MA Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291, 291 (2021).

[16] KAYAK. 2016. *KAYAK + Facebook Messenger.* Retrieved June 5, 2021 from https://www.kayak.com/news/kayak-facebook-messenger/

[17] Philipp Kirschthaler, Martin Porcheron, and Joel E Fischer. 2020. What Can I Say? Effects of Discoverability in VUIs on Task Performance and User Experience. In *Proceedings of the 2nd Conference on Conversational User Interfaces.* 1–9.

[18] Knut Kvale, Olav Alexander Sell, Stig Hodnebrog, and Asbjørn Følstad. 2019. Improving Conversations: Lessons Learnt from Manual Analysis of Chatbot Dialogues. In *International Workshop on Chatbot Research and Design.* Springer, 187–200.

[19] Raina Langevin, Ross J Lordon, Thi Avrahami, Benjamin R Cowan, Tad Hirsch, and Gary Hsieh. 2021. Heuristic Evaluation of Conversational Agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–15.

[20] SeoYoung Lee and Junho Choi. 2017. Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity. *International Journal of Human-Computer Studies* 103 (2017), 95–105.

[21] Chi-Hsun Li, Su-Fang Yeh, Tang-Jie Chang, Meng-Hsuan Tsai, Ken Chen, and Yung-Ju Chang. 2020. A Conversation Analysis of Non-Progress and Coping Strategies with a Banking Task-Oriented Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–12.

[22] Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-end task-completion neural dialogue systems. *arXiv preprint arXiv:1703.01008* (2017).

[23] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 2119–2128.

[24] Chelsea M Myers, Anushay Furqan, and Jichen Zhu. 2019. The impact of user characteristics and preferences on performance with an unfamiliar voice user interface. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–9.

[25] Jakob Nielsen. 1994. *How to Conduct a Heuristic Evaluation.* Retrieved June 10, 2021 from https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/

[26] Jakob Nielsen. 2010. *Mental Models.* Retrieved January 16, 2021 from https://www.nngroup.com/articles/mental-models/

[27] Jakob Nielsen. 2011. *Workflow Expectations: Presenting Steps at the Right Time.* Retrieved January 16, 2021 from https://www.nngroup.com/articles/workflow-expectations/

[28] Rifat Rahman, Md Rishadur Rahman, Nafis Irtiza Tripto, Mohammed Eunus Ali, Sajid Hasan Apon, and Rifat Shahriyar. 2021. AdolescentBot: Understanding Opportunities for Chatbots in Combating Adolescent Sexual and Reproductive Health Problems in Bangladesh. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–15.

[29] Alexander Renkl. 2014. Toward an instructionally oriented theory of example-based learning. *Cognitive science* 38, 1 (2014), 1–37.

[30] Ruhi Sarikaya. 2017. The technology behind personal digital assistants: An overview of the system architecture and key components. *IEEE Signal Processing Magazine* 34, 1 (2017), 67–81.

[31] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. The social roles of bots: evaluating impact of bots on discussions in online communities. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–29.

[32] Arjun Srinivasan, Mira Dontcheva, Eytan Adar, and Seth Walker. 2019. Discovering natural language commands in multimodal interfaces. In *Proceedings of the 24th International Conference on Intelligent User Interfaces.* 661–672.

[33] Justin D Weisz, Mohit Jain, Narendra Nath Joshi, James Johnson, and Ingrid Lange. 2019. BigBlueBot: teaching strategies for successful human-agent interactions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces.* 448–459.

[34] Sihan Yuan, Birgit Brüggemeier, Stefan Hillmann, and Thilo Michael. 2020. User Preference and Categories for Error Responses in Conversational User Interfaces. In *Proceedings of the 2nd Conference on Conversational User Interfaces.* 1–8.