When There is No Progress with a Task-Oriented Chatbot: A Conversation Analysis

Chi-Hsun Li
National Chiao Tung University
Hsinchu, Taiwan
funing314.iem05g@g2.nctu.edu.tw

Ken Chen
National Chiao Tung University
Hsinchu, Taiwan
kent31.iem05g@nctu.edu.tw

Yung-Ju Chang
National Chiao Tung University
Hsinchu, Taiwan
armuro@cs.nctu.edu.tw

MobileHCI '19, October 1–4, 2019, Taipei, Taiwan
 ACM ISBN 978-1-4503-6825-4/19/10…$15.00
https://doi.org/10.1145/3338286.3344407

Abstract
Task-oriented chatbots are increasingly prevalent in our daily life. Research effort has been devoted to advancing our understanding of users' interaction with conversational agents, including conversation breakdowns. However, most research attempts were limited to obversions from a relatively short duration of user interaction with chatbots, where users were aware of being studied. In this study, we conducted a conversation analysis on a three-month conversation log of users conversing with a chatbot of a banking institution. The log consisted of 1,837 users' conversations with this chatbot with 19,449 message exchanges. From this analysis, we show that users more often failed to make a progress in a conversation when they requested information than when they provided information. Furthermore, we uncovered five kinds of intention gaps unexpected to the chatbot, and five major behaviors users adopted to cope with non-progress.

Author Keywords
Chatbot; human-agent interaction, interaction gaps, coping strategy

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

Introduction

Numerous researches have studied users' interaction with intelligent assistants, speech-based CAs, text-based chatbot, virtual assistants, smart home systems, intelligent wrist band, and so on. Chatbot, a text messaging-based conversational agent, is also gaining an increased exposure. A wide variety of chatbots have been deployed on online platforms in different domains [3], such as news, travel, shopping, booking, chit-chat, game, information search, recommendation system, e-learning, entertainment, online customer service, and so on. In particular, task-oriented chatbots, in contrast to a chatbot for chatting, mainly serve to help users perform a specific task in a specific domain. They have been used widely because it can potentially reduce a substantial amount of human labor for customer support if it can resolve relatively simple, but repeated, requests from users. However, if the chatbot cannot handle users' conversations easily, conversation breakdowns may potentially trigger the users to abandon the service [2].

Many researchers have been devoted into improving techniques for humans to more smoothly and naturally interact with conversational agents, for example, via developing better natural language processing techniques or reducing automatic speech recognition errors [9]. Other researchers also aim to advance the understanding of the interactions between users and CAs. For example, [5] attempted to understand how users reformulate their information requests to conversation agents. Other researches studied user's behaviors, expectations, and motivations of using CAs [10, 12]. In particular, a recent study [4] suggests that the motivation for and the type of using CAs affects users' expectation of CAs, that users assess system intelligence, and that the reliability of CAs affect user engagement and ongoing use.

Nowadays, it is still difficult for a chatbot to handle all of the complexities of natural language interactions [11]. As a result, conversation breakdowns are still expected during conversations. While users may prefer the chatbot to repair conversation breakdowns in certain ways [2], users may have their own strategy to cope with a conversation breakdown, which, unfortunately, may include leaving the conversation or even abandoning the entire service. These behaviors are undoubtedly harmful to task-oriented chatbots, which are purposed for fulfilling users' needs, since "leaving" may indicate losing users from the service. These behaviors, however, may be possibly anticipated, and even prevented, if the gaps between users and chatbots can be learned from conversation history.

To uncover these intention gaps, we collaborated with a banking institution and obtained a three-month conversation log of their customers and their chatbot from May 1st 2017 to July 31st 2017. We conducted a conversation analysis on this log, with a total of 1,837 users' conversations, and 19,449 conversation exchanges being analyzed.

In this paper, we present preliminary results from this analysis, where we show that users more often failed to make progress when they requested information than when they provided information. Furthermore, we uncovered five kinds of intention gaps unexpected to

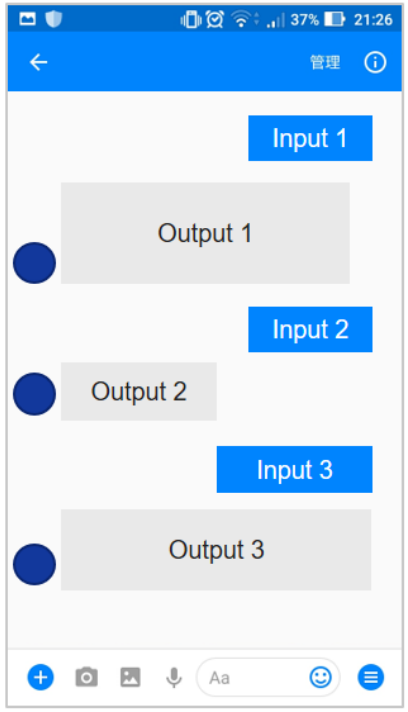

Figure 1: An illustration of a conversation exchange in a user-chatbot interaction from the user's perspective. One exchange means one user's input followed by one chatbot's output. In this example, there are three exchanges.

the chatbot, and five major behaviors users adopted to cope with non-progress, ordered by frequency: including quitting the topic, rephrasing, adding words, repeating, and removing words. While these strategies have been mentioned in previous studies, we reveal the proportion of them in a real usage of a task-oriented chatbot in users' daily lives.

Method

Data Collection and Processing

As mentioned above, we collected a three-month conversation log of users and a task-oriented chatbot from a banking institution. The conversation log recorded its customers' conversations with the chatbot from May 1st 2017 to July 31st 2017. The chatbot used the Facebook Messenger chatbot service, of which the services included currency, credit card, house loan, bank account, investment, etc.

The original dataset contained 2,597 users' conversations and 24,074 exchanges (i.e. one user's input followed by one chatbot's output, see Figure 1). The data were stored in a spreadsheet file, with each row representing a conversation exchange (see Figure 2). The log only contained the text output of the chatbot but did not contain information about the visual elements that users saw in a chat window. In addition, the input text did not distinguish between the text users typed by themselves and the response button that users chose. Since users react to what they see and what is available, it is important to know what exactly they would see during the conversation. Therefore, the researchers first interacted with the chatbot following the same flow of the recorded conversation exchanges and added more details about each conversation exchange. This replication process took a substantial amount of time, but it is critical to do so.

ChannelID	Input	Intent	Output	Time
User1	Input1	Intent1	Output1	20JUN2017:19:15:47.833
User1	Input2	Intent1	Output2	20JUN2017:19:15:55.497
User1	Input3	Intent3	Output3	20JUN2017:19:16:07.993

Figure 2: original data form

ChannelID	Input	Intent	Output text	Card	Button	Quick response button	Time
User1	Input text1	Intent1	Output text1				20JUN2017:19:15:47.833
User1	Input text2	Intent1	Output text2				20JUN2017:19:15:55.497
User1	Input text3	Intent3	Output text3				20JUN2017:19:16:07.993

Figure 3: data form after transcribed

Eventually, each conversation exchange had the attributes of the *input text, output text, card, button, quick response button* (see Figure 3, 4 & 5). We then removed two kinds of exchanges not suitable for analysis. The first kind was exchanged that we could not replicate because the output from the chatbot had changed (e.g. seasonal promotions from the institution that were no longer available when we obtained the data). The other kind was exchanged occurring at the end of the log, of which we did not know the final outcome of the conversation. After this data cleaning process, we removed 760 users' conversations and 4,125 exchanges, resulting in 1,837 users' conversation and 19,449 exchanges in the final dataset. Among these exchanges, 39.8% of the inputs were users typing their own words, with an average length of 5.78

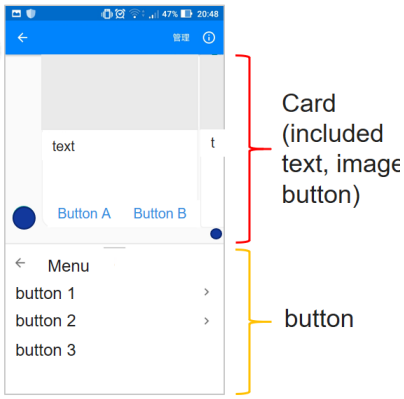

Figure 4: examples of the chatbot's display

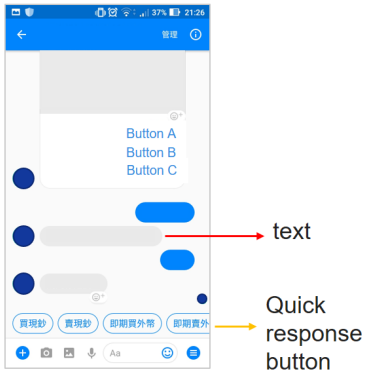

Figure 5: examples of the chatbot's display

words (SD=7.25); 19.8% were clicking buttons (see Figure 4). And the rest of the 40.4% were either users clicking on quick response buttons (see Figure 5) or themselves typing the same content appearing on the quick response buttons, which we could not distinguish.

Conversation Analysis

We used conversation analysis, an inductive process for analyzing how human interaction is organized into sequences of action and systematic practices [1] on the conversation log. e focused on turn-taking, structural organization (opening-request-advice/information-closing) [6], sequence organization (request-link-thanks), and lexical choice [12] in the conversations. We examined how sequences of exchanges were grouped and situated in particular instances of conversation line by line, and assigned codes to each individual exchange. This allowed us to organize interactions into sequences of actions, which is a key element of conversation analysis [1]. The coding process was inductive and data-driven. The first author of the paper generated the first set of 106 codes. The codebooks contained six main categories: Event topic (5), TimeBreak (3), Session (4), User Input (43), Chatbot Output (24), Next Step Behaviors (27). Then, we recruited a second coder to join the coding process to ensure the reliability of the codebook. The two coders used 2% of the full dataset to apply the original codes. They iteratively discussed and revised the codes, with a third researcher joining the discussion weekly about the high-level themes. This process continued until both coders' codes agreed with each other entirely. The two coders then test their inter-coder reliability with a representative sample of the data (10%). Through the same iterative process of discussion and generating, revising, removing, and combining codes, the two coders reached a Cohen Kappa of 0.802, indicating high reliability between the two coders [7]. The coders then started coding the rest of the dataset. The final codebook contained 88 codes: Event topic (5), TimeBreak (5), Session (4), User Input (34), Chatbot Output (27), Next Step Behaviors (13). Each conversation exchange was assigned these 88 codes.

Preliminary Results

Users' Interaction Characteristics with the Chatbot

The first result is regarding the types of user interaction with the chatbot. All the exchanges were classified by two dimensions: 1) whether the user input was intended, and 2) whether the user input made a progress. From Figure 6, we see that information provision (56%, N=10,931) was the top type of exchange in the conversation log, followed by user request (38%, N=7,350), and chat/noise (6%, 1,156). Despite a very small portion of chat/noise, the result suggests that users indeed attempted to casually chat with this chatbot, even though it was designed to answer banking-related information tasks.

What Types of Exchange More Often Led to Progress?

Among all 19,449 exchanges, 88.1% made progress, and 11.9% did not. In particular, users the most often made progress when the exchange was information provision (98.1%), followed by user request (80.1%) and chat/noise (42.7%). In other words, users were more likely to proceed when they provided information than they requested information. This implies that the conversation breakdowns between users and the chatbot more often occurred when users asked than when users answered. Our observation was that in asking questions users more often used their own way to ask questions, especially when they were not prompted what kind of questions chatbot could answer.

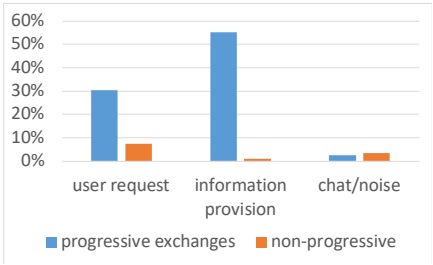

Figure 6: percentage of basic user interaction

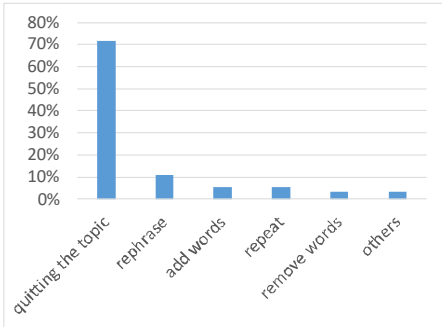

Figure 8: percentage of the way users cope the gaps

Six types of cause of non-progress	Mis-recognize	Could not recognize
intended interaction	43.1%	46.4%
extra explanation	1.6%	2.2%
unexpected restart	0.3%	0.3%
unexpected return	0.4%	1.3%
unfinished sentence	0.6%	2.2%
finishing the unfinished sentence	0.8%	1.0%

Table 1: type of interactions gaps and percentage of non-progressive exchanges

In contrast, when being prompted to provide information, usually the instruction was more explicit for users to follow.

What were the Sources of Non-Progress?
There were two kinds of situation in which the conversation did not make progress. The first was when the chatbot misunderstood the user's intended meaning (coded as *mis-recognize* in Table 1). The other was that the chatbot was not able to recognize the user's intended meaning (coded as *could not recognize* in Table 1). For both types of non-progress, we identified six types of causes, including one expected intention, and five types of intention gaps. Most of the non-progressive exchanges were expected intentions (89.5%), as shown in Table 1. In other words, the texts they enter was within the range of the service of the chatbot; however, the chatbot simply misunderstood the intention or was unable to recognize the intention. It was because users often used colloquial and local language way to ask or to provide information as they talk to people, which was not correctly recognized by the chatbot. However, the rest of five intention gaps were what we found interesting: extra explanation (3.7%), unexpected restart (0.6%), unexpected return (1.7%), unfinished sentence (2.8%), and finishing the unfinished sentence (1.7%). Although these five unexpected intention gaps only contributed nearly 10% of the non-progress exchanges, the portion is still non-trivial if they ultimately led to task abortion. Note that some of these were common in users' daily texting, such as adding extra explanation not finishing a complete statement. However, they are challenging for a chatbot to handle because an assumption that an intention of the user can be completely conveyed via a sentence was violated.

How Did Users Cope with these Intention gaps?
Finally, there were two major kinds of user behaviors when users did not make progress. The first was users trying to quit the topic (71.5%), such as asking a new question, leaving the conversation, or turning to chat. The other was retrying on the same topic (28.5%), including rephrased the words (11%), adding words (5.5%), repeating with the same words (5.2%), removing words (3.4%), and others (3.4%). The high frequency that users would give up the topic suggests that conversation breakdowns are important to solve.

Future Work

We present the first analysis on a natural conversation of users with a task-oriented chatbot via a conversation analysis on a three-month conversation log. We show that intention gaps between the two entities more often

occurred when users requested information than when they provided information. Although the majority of non-progress was due to the misunderstanding and the failure of recognition on users' input of which the intention was expected, there were roughly 10% of user input that belonged to unexpected intentions to the chatbot. Interestingly, some of these *intention gaps* were common in our daily life conversations. Knowing the presence of these behaviors allows us to anticipate them and prepare for them in the future. Finally, we show that users more often quitted the topic than retrying when they encountered non-progress.

References

1. Paul M. Aoki , Margaret H. Szymanski , Luke Plurkowski , James D. Thornton , Allison Woodruff , Weilie Yi, 2006. Where's the "party" in "multi-party"?: analyzing the structure of small-group sociable talk. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, November 04-08, 2006.
2. Zahra Ashktorab, Mohit Jain, Q. Vera Liao, & Justin D. Weisz. 2019. Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), 1–1
3. Jonathan Grudin, Richard Jacques. 2019. Chatbots, Humbots, and the Quest for Artificial General Intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* (CHI '19), 1-11.
4. Mohit Jain, Ramachandra Kota, Pratyush Kumar, & Shwetak Patel. (2018). Convey: Exploring the Use of a Context View for Chatbots. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), 1–6.
5. Jiepu Jiang, Wei Jeng, & Daqing He. 2013. How Do Users Respond to Voice Input Errors?: Lexical and Phonetic Query Reformulation in Voice Search. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 143–152.
6. Jane Lockwood. 2017. An analysis of web-chat in an outsourced customer service account in the Philippines, *English for Specific Purposes*, 47, 26-39.
7. Matthew Lombard, Jennifer Snyder-Duch, Cheryl Campanella Bracken. 2010. Practical Resources for Assessing and Reporting Intercoder Reliability in Content Analysis Research Projects.
8. Ewa Luger, Abigail Sellen. 2016. "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* (CHI '16), 5286-5297.
9. Joanne Meredith. 2017. Analysing technological affordances of online interactions using conversation analysis. Journal of Pragmatics, 115, 42–55.
10. Vera Q. Liao, Muhammed Masud Hussain, Praveen Chandar, Matthew Davis, Marco Crasso, Dakuo Wang, Michael Muller, Sadat N. Shami, and Werner Geyer. 2018. All Work and no Play? Conversations with a Question-and-Answer Chatbot in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18)
11. Filip Radlinski, Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (CHIR '17), 117-126.
12. Wyke Stommel, Trena M. Paulus, David P. Atkins, "Here's the link: Hyperlinking in service-focused chat interaction," *Journal of Pragmatic*s 115 (2017) pp.56-57.