Position paper presented at CONVERSATIONS 2022 – the 6th International Workshop on Chatbot Research, Applications and Design, hosted by the University of Amsterdam, Netherlands, November 22-23.

Conversation Mining for Customer Service Chatbots

Daniel Schloß[1][0000-0001-9158-3528] and Ullrich Gnewuch[2][0000-0003-1423-1777]

[1] Karlsruhe Institute of Technology, Kaiserstraße 12, 76131 Karlsruhe, Germany
[2] Karlsruhe Institute of Technology, Kaiserstraße 12, 76131 Karlsruhe, Germany

Abstract. More and more companies are using chatbots in customer service. The large number of chatbots and their interactions with customers produce a huge amount of data, which is useful to track the usage and performance of the chatbot. However, many established performance metrics (e.g., intent scores, conversations per day) could be considered too intuitive to be helpful and are either at a very high level or at the level of single question-answer pairs. Our research aims to address this challenge by presenting a novel approach and system for conversation analysis of customer service chatbots. More specifically, we extend established metrics and concepts with ideas from process mining since every conversation with customer service chatbots can be interpreted as a sequence of discrete steps. This paper presents the methodological foundations for our approach, which we call *conversation mining*, and demonstrates its potential with first insights into our prototype. Ultimately, we aim to draw the attention of chatbot researchers and practitioners to the value of conversation data by describing a novel approach for automatically processing and analyzing at a process level.

Keywords: Chatbots, Customer Service, Data Analytics, Conversation Mining.

1 Introduction

In recent years, especially since 2016, chatbots have become increasingly widespread [5]. They automate conversations and the handling of business processes and are particularly suitable for automated responses to simple customer inquiries or for automating clearly defined standard processes. Chatbots therefore are ideal and also popular as an additional channel for customer service, since many similar concerns that need to be processed efficiently occur there [6]. The strengths of chatbots lie in the permanent availability and the possible parallelization when processing requests [4]. Since chatbots therefore often handle high request volumes, customer service managers need analytical tools to be able to monitor the performance of chatbots as easily as possible [2]. Systems for chatbot analytics are also essential because chatbots operate independently. Either chatbots complete customer touchpoints or there are dropouts somewhere in the service process [8]. Thus, to monitor the bot's work and also to prevent "bot rot," i.e., neglected chatbot maintenance that impairs the customer experience, customer service managers need both, the conversation data and systems to analyze it efficiently. With the support of such systems, they can make informed decisions on the design of the chatbot and leverage the huge potential hidden in the numerous individual user experiences that have been collected [2, 11, 19].

Because of that importance, chatbot research increasingly focused on the usage data that chatbots produce in recent years. For example, it has been investigated how breakdowns in dialogs can be detected (and prevented) [1, 3, 11] or which concepts guide the analysis of conversational log data, e.g. with regard to user experience [2, 7, 19]. Since users often leave the chatbot when there are problems in the conversation [11], the investigation of dialog situations is very important. Although this can be done via manual review and coding [10], scalable approaches for large volumes of conversations are needed due to efficiency reasons. Furthermore, when investigating chatbot conversations it is important that not only individual question-answer pairs, so-called turns, which many research projects addressed, are evaluated, but also dialogs in their entirety [2, 19]. We aim to address this existing research gap for the automated analysis of entire user-chatbot dialogs [12, 19].

Since typical task-oriented chatbots in customer service still rely on buttons, click paths and/or intent recognition instead of self-learning techniques, human judgment and decision-making is required for their improvement [7, 8, 10]. Thus, we aim to support customer service managers with actionable insights to easily identify where to improve the topics (1), the natural language understanding (NLU) training (2) or the dialog design of a chatbot (3). To this end, we propose a conversation mining system that uses established performance metrics and extends them with an established visual approach borrowed from the field of process mining, which we introduce in the following [18, 19]. The conversation mining project is being conducted in collaboration with a chatbot development company that provides customer service chatbots for the energy industry.

2 Related Work

2.1 Chatbots in Customer Service

When a customer wants to handle a service request through the chatbot channel, for example changing his/her contact information, it is important that no gap occurs between his/her expectation and the perceived/actual service performance [9]. For this to be guaranteed, two conditions must be met. First, the concern the concern the customer expresses to the chatbot must be thematically present in the chatbot and trained in the NLU so it can be understood and processed. Second, the process of handling the concern must meet the customer's expectation. Some chatbots do not resolve concerns in the channel but only direct customers to the right destination (link). Other chatbots can send data inputs from the frontend to ERP systems or retrieve data from them [7]. In general, problems within the service processes applied in chatbot dialogs arise when the assumptions of the service provider deviated from the expectations the customers have [9]. Examples of this are: (1) Customers ask a chatbot about unfamiliar topics, such as special products [7]. (2) The chatbot does not understand the customer, for example because unknown synonyms are used [13]. Or (3) a subsequent service process represented in a fixed chatbot dialog confuses or overwhelms customers, for example because they don't know whether they have to enter the customer number in XX or X-X format [19]. Formats like these and dialog or query logic can be determined by connected ERP systems, their data structure and web services [19].

2.2 Analytics and Process Mining for Chatbot Conversations

Analyses of customer chatbot conversations show where assumptions about customer behavior and actual behavior form a gap [9]. The first fundamental problem, (1) missing topics in the chatbot, can only be analyzed indirectly in an automated way [1, 7, 10]. Since chatbots do not know what they do not know, “unknown topics” are not logged. Requests for unknown or out-of-scope topics will instead end up as misrecognized intents (mismatches) or unrecognized intents (low scores) in the log data [10]. While low intent scores, also useful to investigate (2) NLU errors on in-scope topics, can be easily filtered out via chatbot analytics systems, mismatches are difficult to identify and would require human resources [10, 13]. What can be analyzed, however, is the subsequent course of the conversation: For example, if a customer was not correctly understood, he may a) get angry in the following utterance, b) repeat or paraphrase, or c) abandon the attempt [2, 11]. This shows that the detection of problematic conversation situations in retrieval chatbots does not only take place on the single-row-level of an utterance and its intent classification, but that the data analysis has to be extended to the “vertical”. In some cases (e.g., mismatch), the cause of the error is not automatically identifiable, but error indicators (e.g., negative customer sentiment) in the progress of the conversation are. In other cases, the progress of the conversation is inconspicuous, but there were detectable errors (e.g., technical error in the chatbot response). Thus, for automated conversation mining, customizable filter functions that target the occurrence of errors and error indicators in the entire conversation are helpful [2, 15, 19].

If the customer's topic selection and the intention classification worked, at last (3) there may occur problems in the service processes triggered by the chatbot. Whereas customers in an “open” conversation situation have the opportunity to perform a variety of actions or inputs, the recognition of some intents may also be followed by a static dialog, for example, because data must be queried from the customer and validated [15, 18]. In this case, the chatbot takes the conversational lead and guides the customer through a process of information exchange, e.g., starting with an identification [14]:

User: “I would like to cancel my flight” *(Cancel_Intent)*
Bot: “You want to cancel your flight? Please tell me your contract number:”

N-step dialogs or processes like this one can be analyzed using process mining methods. When a process requires multiple user inputs, e.g using forms or single texts, service managers had an expectation/idea of how customers would behave and how the dialog should be designed to be customer-friendly and smooth. This expectation can be compared aggregated with the actual behavior of the customers visible in the log data (Conformance Checking [18]). Hence, there is *1* ideal variant as well as *m* real variants of how customers can move through an *n*-step process. In practical terms, the number of variants *m* depends on the customer's “degrees of freedom” on the interface, for example, whether the customer is allowed to click a button even though he is asked for input. Moreover, if logging is insufficient, the logged variants do not reflect the real variance in customers' behavior (interactions) [18]. However, assuming the data source is rich, there is a lot of potential in process mining to check for customer-service provider expectation gaps in static dialogs [9, 15, 18]. Furthermore, process mining analysis can also be conducted for the non-chatbot-guided conversation parts, illustrating (with a high number of variants) typical paths as users "move" through the chatbot. A simple application of this is to calculate and visualize the most frequent consecutive intents (e.g., "60% of customers move from Intent B to Intent A", Process Discovery [16, 18]).

3 Method: Chatbot Conversation Mining

3.1 Starting Point and Definition of Objectives

Starting point of the ongoing research project were six 1.5-hour interview sessions with four chatbot developers (Product Owner, Operations, NLU, UI/UX) of our industry partner and two customer service managers of a B2B customer. All of them emphasized the relevance of frequent and automated monitoring and analysis of the chatbot's performance. When asked about possible reasons for problems in customer chatbot conversations, analogous to chapter 2, response content, NLU problems and errors in bot-guided dialogs were described as sources of errors besides user-related errors and a weak expectation management. Since the experts especially demanded analytics related to the progress of conversations (e.g., to see what customers do before they click a specific link), in addition to simpler conventional metrics, e.g. frequency of conversations, it was decided to deploy process mining technicques as central idea of the system.

3.2 Concept and Development

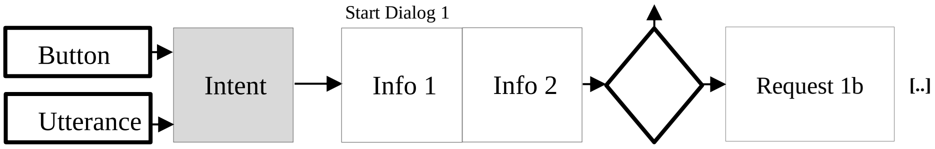
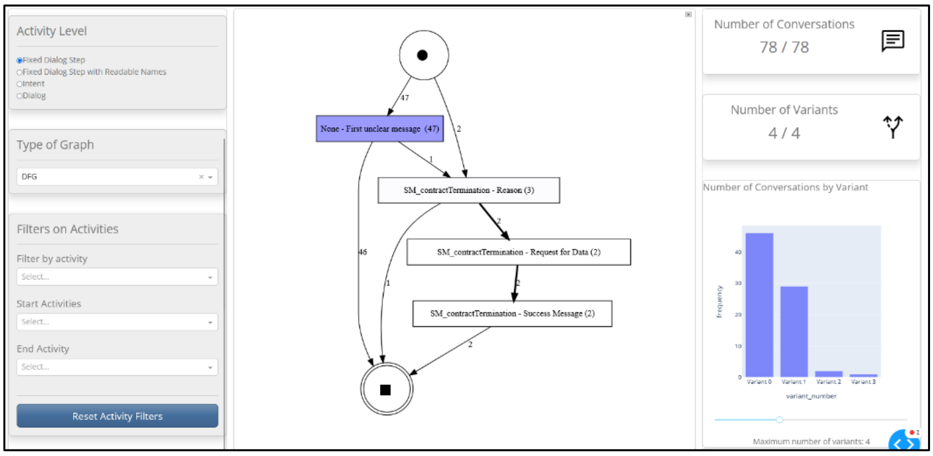


Fig. 1. Exemplary depiction of a control-flow diagram of a chatbot dialog

The development of the conversation mining began with capturing the technical preconditions. The central chatbot application of our industry partner is based on the Microsoft Bot Framework and already records conversational data in a SQL-database [14]. We checked the data basis for richness and quality and found the necessary prerequisites in the raw data for simple functions such as data filters like uniquely assigned and logged conversation IDs. In addition, classification data from the NLU is logged. To plan the single steps for the process mining approach, we started looking at a mostly static dialog that was especially problematic for the chatbot users. We drew a control-flow diagram, shown exemplary in **Figure 1**, consisting of all possible chatbot or user activities (e.g., the chatbot starting Dialog 1). Including user actions as well as data on the internal chatbot dialog management, it provides an overview on all the dialog paths that a user can follow in the interaction [18]. The control-flow diagram shows the options of a customer when he or she is in a guided dialog situation. When we compared this model of possible chatbot and user activities with the actual log data we found small discrepancies regarding events that were not sufficienctly logged or not differentiated [18]. For example, link clicks were only logged at the conversation level, and for bot responses logged as text, there was no identifier for whether it was an arbitrary 1-message response or a specific step in a guided dialog. Therefore, for the prototype, we used a mapping between the logged texts of the chatbot and the texts maintained in the chatbot response database to subsequently reproduce dialog positions. In addition, we defined all the requirements for raw data logging to meet the principal requirements of a process mining event log, such as distinctly logging each real-world event [18].

4 Conversation Mining Prototype



Activity Level

● Fixed Dialog Step
○ Fixed Dialog Step with Readable Names
○ Intent
○ Dialog

Type of Graph

DFG

Filters on Activities

Filter by activity: Select...

Start Activities: Select...

End Activity: Select...

Number of Conversations: 78 / 78

Number of Variants: 4 / 4

Maximum number of variants: 4

Reset Activity Filters
47
2
None - First unclear message (47)
1
SM_contractTermination - Reason (3)
2
SM_contractTermination - Request for Data (2)
46
1
2
SM_contractTermination - Success Message (2)
2
Number of Conversations by Variant
frequency
0
10
20
30
40
Variant 0
Variant 1
Variant 2
Variant 3
variant_number

Fig. 2. Prototype with Filters, Process Visualization and Basic Metrics

As a preliminary result, we developed an initial conversation mining prototype shown in **Figure 2**. The system helps to filter and analyze conversations according to quality and performance criteria. It allows to select conversations with certain activities, for example, those in which an error occurred. At the same time, it visualizes the paths and their respective utilization across all conversations. For this purpose, the rectangles sum up the cases (conversations) in which certain activities occurred and on the arrows the number of the respective transitions can be seen. For our first examined dialog, we were able to determine which options or paths were hardly used at all by analyzing the numbers of conversations reaching certain bot answers/dialog steps. The prototype also allows to identify problematic situations in the form of loops (an arrow goes from a rectangle to the same one) or terminations (direct arrows from a rectangle to the end node). It can be used by customer service managers to explore the sequences of topics in the chatbot or to identify how extensively particular dialogs are being used. The prototype thus makes topics, utilization, problems or terminations transparent, but is of course dependent on the data basis, the event log. The analyses can serve as a basis for decisions on dialog design, topics, and to address particularly critical dropout situations.

5 Conclusion and Future Work

As we were able to show in this position paper, there is great potential in the analysis of customer service chatbot log data. In particular, if whole static processes/dialogs or conversations are investigated, this opens further opportunities for customer service operators to identify chatbot weaknesses and initiate improvements [2, 18]. With our conversation mining system as well as the associated design knowledge we therefore contribute to a stronger understanding of user-chatbot interactions. We plan to finalize the system with further filter options and improved event logs [18]. A limitation of our work is that our current process-based analysis is taylored to text-based chatbots. Future research could build on that and transfer the concept to other conversational channels.

References

1. Akhtar, M., Neidhardt, J., and Werthner, H. The potential of chatbots: Analysis of chatbot conversations, Proceedings - 21st IEEE Conference on Business Informatics, 1, pp. 397–404. (2019).
2. Beaver, I., and Mueen, A.. Automated conversation review to surface virtual assistant misunderstandings: Reducing cost and increasing privacy, 34th AAAI Conference on Artificial Intelligence, pp. 13140–13147. (2020).
3. Benner, D., Elshan, E., Schöbel, S., and Janson, A. What do you mean? A review on recovery strategies to overcome conversational breakdowns of conversational agents, ICIS 2021 Proceedings. (2021).
4. Brandtzaeg, P. B., and Følstad, A.. Why people use chatbots, Lecture Notes in Computer Science, 10673 LNCS, 377–392. (2017).
5. Dale, Robert. The return of the chatbots. *Natural Language Eng.* 22.5, 811-817 (2016)
6. Følstad, A., and Skjuve, M. Chatbots for customer service: user experience and motivation, CUI '19: Proceedings of the 1st International Conference on Conversational User Interfaces, pp. 1–9. (2019).
7. Følstad, A., and Taylor, C. Investigating the user experience of customer service chatbot interaction: framework for qualitative analysis of chatbot dialogues, Quality and User Experience (6:1), pp. 1–17. (2021).
8. Grudin, J., and Jacques, R. Chatbots, humbots, and the quest for artificial general intelligence, Conference on Human Factors in Comp. Systems - Proceedings, pp. 1–11. (2019)
9. Halvorsrud, R., Kvale, K., & Følstad, A. Improving service quality through customer journey analysis. Journal of service theory and practice. (2016).
10. Kvale, K., Sell, O. A., Hodnebrog, S., and Følstad, A. Improving conversations: lessons learnt from manual analysis of chatbot dialogues, Lecture Notes in Computer Science, 11970 , pp. 187–200. (2020).
11. Li, C. H., Yeh, S. F., Chang, T. J., Tsai, M. H., Chen, K., and Chang, Y. J. A Conversation Analysis of Non-Progress and Coping Strategies with a Banking Task-Oriented Chatbot, Conference on Human Factors in Computing Systems - Proceedings, pp. 1–12. (2020).
12. Przegalinska, A., Ciechanowski, L., Stroz, A., Gloor, P., and Mazurek, G. In bot we trust: A new methodology of chatbot performance measures, Business Horizons (62:6), pp. 785 – 797. (2019).
13. Reinkemeier, F. and Gnewuch, U. Designing Effective Conversational Repair Strategies for Chatbots, ECIS 2022 Research Papers (2022).

14. Rozga, S. Practical bot development: Designing and building bots with Node. js and microsoft bot framework. Apress. (2018).
15. Schloss, Daniel, Ulrich Gnewuch, and Alexander Maedche. Towards Designing a Conversation Mining System for Customer Service Chatbots, ICIS 2022 Research Papers (2022).
16. Topol, Zvi, Using Process Mining to Improve Conversational Interfaces", https://fluxicon.com/camp/2019/3, last accessed 2022/25/09.
17. Van Der Aalst, W. et al. Process mining manifesto. International conference on business process management. Springer, Berlin, Heidelberg, (2011).
18. Van Der Aalst, W. Process mining: Overview and opportunities. ACM Transactions on Management Information Systems (TMIS), 3(2), 1-17. (2012).
19. Yaeli, A., and Zeltyn, S. Where and Why is My Bot Failing? A Visual Analytics Approach for Investigating Failures in Chatbot Conversation Flows, Proceedings - 2021 IEEE Visualization Conference - Short Papers, VIS 2021, pp. 141–145. (2021).