Comparing Moderation Strategies in Group Chats with Multi-User Chatbots

Nicolas Wagner
Ulm University
Ulm, Germany
nicolas.wagner@uni-ulm.de

Matthias Kraus
Ulm University
Ulm, Germany
matthias.kraus@uni-ulm.de

Tibor Tonn
Ulm University
Ulm, Germany
tibor.tonn@uni-ulm.de

Wolfgang Minker
Ulm University
Ulm, Germany
wolfgang.minker@uni-ulm.de

ABSTRACT

The increasing capabilities of chatbots will become more and more important in the near future. In this paper, we introduce a conversational system which connects a chatbot with a mainstream messaging service in a multi-user scenario. While there are already numerous options for single-user bots, ready-to-use systems for multiple users and group chats are scarce. This work thus aims to get insight into how such a group chatbot should behave during a multi-turn conversation. For this, we implemented and evaluated four different moderation strategies in an everyday use-case: the planning and negotiation of a joint appointment. In our subsequent user study with 40 participants, we investigated how the different strategies were perceived and what influence they had on the acceptance, usability and efficiency of the system. Our evaluation results show that users' perceptions of innovation and inventiveness of the bot were influenced by the moderation strategies.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **HCI design and evaluation methods**; **User studies**; *Collaborative interaction.*

KEYWORDS

human-machine interaction, multi-user dialogue systems, user centered evaluation, multi-user chatbots

ACM Reference Format:
Nicolas Wagner, Matthias Kraus, Tibor Tonn, and Wolfgang Minker. 2022. Comparing Moderation Strategies in Group Chats with Multi-User Chatbots. In *4th Conference on Conversational User Interfaces (CUI 2022), July 26–28, 2022, Glasgow, United Kingdom.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3543829.3544527

CUI 2022, July 26–28, 2022, Glasgow, United Kingdom

ACM ISBN 978-1-4503-9739-1/22/07.
https://doi.org/10.1145/3543829.3544527

1 INTRODUCTION

It is hard to imagine life today without the use of messaging applications on smartphones. An entire range of such services are available for different platforms, most notably WhatsApp, Facebook Messenger, Signal, and Telegram[1]. With a total of roughly 3 billion users already, the numbers are expected to rise to 3.5 billion users in the next 3 years [8]. As a result, even a substantial proportion of non-technological affine people have learned to exchange messages either in private or in group chats. Being originally developed for human-human interaction, some of these apps already allow users to communicate with conversational agents.

Especially the times of the Corona pandemic and its concomitant social distancing and teleworking have shown that it is elementary for teams to be able to collaborate without face-to-face contact. Although users are accustomed to group chats, it is still uncommon to have a conversational agent as a partner in that group. While there are a large and growing number of chatbots designed to interact with a single user, research also needs to address the issue and challenges of group conversations. Thus, this work presents an approach towards assisting an appointment negotiation in a multi-user setting through intelligent support by a chatbot. According to McTear [6], reasons for relying on chatbots may be a low barrier to entry for users and a more intuitive way of interaction with services, resources, and data. Group chatbots could lead to a new way of conversation with virtual systems and make communication easier, faster, and more efficient.

To explore how such a conversational group agent should participate in or lead a discussion, we implemented and evaluated four different strategies for the system's moderating behaviour in a realistic scenario. Here, different patterns were employed to determine *how* and *when* the chatbot is supposed to engage in the discussion with group partners. For this, we implemented a Dialogue System (DS) which is directly connectable with the messaging service Telegram. We carried out a user study with 40 participants to observe how users perceived the different moderation behaviour during the experiment and to identify the strategy that was rated best in terms of acceptance, usability and efficiency.

The results showed that the behaviour and engagement of the system indeed have an impact on the users' perception of innovation and inventiveness of the conversational assistant. This paper is structured as follows: After discussing related work in the next

[1] https://telegram.org/

section, we present and briefly describe our experimental design in Section 3. Subsequently, we show the results of our user evaluation. Finally, we come to a conclusion in Section 5.

2 RELATED WORK

In recent years, research on chatbots has gained a boost in attention. Some of these systems, in the recent past for example [1, 3], aim to develop a conversational agent that is able to communicate almost like a human being. In contrast, we examine another type of chatbots in our work: a system that supports its users during a goal-directed conversation. For example, the chatbots presented in [2, 10] are able to respond to frequently asked questions or common problems in a special campus environment. However, like the vast majority of published chatbots, they are designed for conversations with a single user. Since it seems impractical to directly transfer single-user chatbot and their conceptual behaviour to a multi-user setting, such systems are not considered further in this paper.

An approach to multi-user chatbots is described in [5], where the authors implemented a goal-directed group chatbot. For our work, we also intend to incorporate a motivational element into system behaviour. Given that we want users to interact with our system in the most natural and realistic way possible, we refrain from imposing a general time limit for our experiments. The multi-user chatbot developed by [11] is designed to schedule tasks in a collaborative team. However, users have no direct contact with each other, as they only experience a one-shot interaction with the bot. Contrarily, our system extends the scheduling service to multi-turn conversations between a mixed human-machine team.

Furthermore, we conducted an analysis of group chat corpora for this work. The work of Seufert et al. [9] presents a related analysis of text messages with a focus on WhatsApp group chats. For our system, it has proven useful to know how and what users normally write in order to adapt the language model and the DS to it. As a consequence, we were able to determine a variety of terms, abbreviations, emojis and other things that are intermittently used in group chats. The actual realisation of our system is described in the next section.

3 EXPERIMENTAL DESIGN

This section presents the architecture of our system and the concept of our domain. We implemented all modules in Python and SQL.

3.1 System Design

For our user interface, a messaging application had to be selected to enable users to interact with the chatbot using text. As mentioned earlier, we decided for the messaging service Telegram. The Telegram framework allows to easily create and operate a new bot via an API. For our DS, we opted to implement the chatbot using the free and open-source Rasa framework[2]. It features components for Natural Language Understanding (NLU) and Dialogue Management. Within the scope of related work, we could not find any published configuration data for the intended domain in a multi-user scenario. Consequently, it was necessary to implement both components completely by ourselves. In addition, several Rasa classes and concepts had to be adapted, as they were not designed for an interaction with multiple users. The interaction strategies were implemented using statistical methods like the Transformer Embedding Dialogue policy [12]. In exceptional cases where this was not possible due to a lack of training data, a decision tree was used. Since our approach does not require any special hardware or software, we hosted the two modules on a virtual machine running on an Apache2 server in a cloud-based setup. Fig. 1 illustrates the overall system architecture.

[2]https://rasa.com

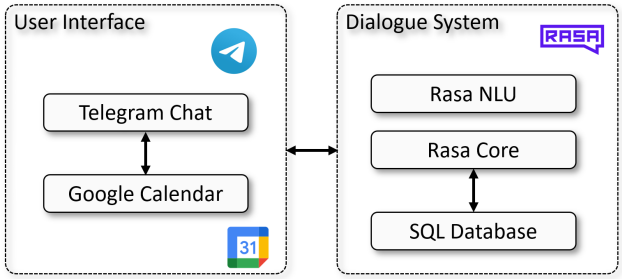

Figure 1: Architecture of the overall system.

3.2 Domain Design

We opted for a simple but common task for groups as our domain: negotiating an appointment. Users are given the task of agreeing on a day, a time, and a meeting point within the next week. They are provided with a specially prepared Google calendar, which already contains a series of appointments, but also holds some unoccupied time slots. Initially, the users are assigned to a group chat together with the bot, and they are told that the system is able to assist them in their negotiation. The users are able to propose specific dates for a meeting. If a proposed date is already blocked for at least one person, the chatbot informs about the scheduling conflict and concurrently suggests a potential time slot. If a user rejects this slot, the system automatically marks it as inadmissible in the calendar. If a team partner does not agree with the time, they have the option to ask the bot for a suggestion for an earlier or later time slot. Unless explicitly requested otherwise, a standard duration of one hour per appointment will be assumed. After agreeing on a time slot, users have to discuss the location of their meeting. For this, the bot is able to offer several suggestions (e.g. gym or restaurant). If one or more users have not written anything for a while, the chatbot will start to motivate them to rejoin the discussion. Once all details have been settled, the appointment is saved in the users' respective calendars. Subsequently, it is possible to schedule another meeting.

An exemplary conversation with could look like this:

U_1: "Hi, how are you?"
U_2: "Hey, thanks I am fine. How about you?"
U_1: "I am fine too :) Would you like to meet on Thursday, let's say 2 pm."
S: "I am sorry, but one of you is already scheduled in another event on Thursday at 2 pm. Alternatively, all team partners would be available on Thursday at 10 am."

U_2: "Hmm, I'm okay with it."
U_1: "Is it possible to meet later?"
S: "Alternatively, all team partners would be available on Thursday at 5 pm."
U_1: "That's fine."
U_2: "Sure!"
S: "I saved the date and time that suits you: Thursday at 5 pm. Where do you want to meet?"
U_2: "I would like to meet at the gym."
U_2: "I am already looking forward to Thursday :)"
S: "Please wait, your colleague has to agree to the location."
U_1: "I agree to the gym."
S: "I saved the location: gym. The appointment is now added to your calendars. Have a great time at your meeting."

The above functionalities are available independently of the applied moderation strategy, but the way the system intervenes in the negotiation process changes. The differences are described below.

3.3 Moderation Strategies

Four different moderation strategies for dialogue management were designed and implemented in our system. The strategies differ in the way the system interacts during the moderation of the negotiation between the users.

(1) **Baseline** This strategy can be considered as baseline system behaviour. Here, if users have a request, they have to explicitly ask for support. This means that the bot will not start interacting with users until someone writes "Help." or something similar.

(2) **Notification** This behaviour triggers a system reaction when any user specifies a date or time. The response is the indication if all users are available at this time or if a person is already occupied. However, the system does not mention which user is busy, nor does it give further information. The bot periodically attempts to obtain the date, time, and preferred location of users, if not already set.

(3) **Private Negotiation** Originally, the Private Negotiation strategy behaves similar to the previous one. Nevertheless, there is an additional function in the case of a scheduling conflict. If a user proposes a certain time and only one of the others already has another appointment there, the bot is able to text a user individually. For this, the system will create a new private chat to the respective person and ask if the event at that time slot should be moved. If the person agrees, the system will shift the dates in the calendar and inform the group that the proposed time slot is available for all users.

(4) **Group Negotiation** As a more persistent version than the Private Negotiation strategy, this behaviour also attempts to resolve a scheduling conflict. If a user is already busy on a suggested time slot, the system posts directly to the group chat that this person is not available and asks if the person would agree to reschedule the event. To avoid confusion, the name of the concerned person will be mentioned in the message. By doing so, we assume that this will put more pressure on users to agree to the rescheduling and to participate in the discussion.

For all strategies, we have implemented a fallback action in case there is a misunderstanding or the user input could not be categorised. The bot then reported that it failed to understand the user and requests a repetition or rephrasing. If users started to talk about off-topic areas, an out-of-scope message was sent reminding them to stay on task. The evaluation procedure and its results are presented in the next section.

4 EVALUATION AND RESULTS

Following the development of the four moderation strategies, the strategies were evaluated as independent variables in a user study. As we aimed to investigate the robustness and effects of the conversational behaviour in this work, we decided on a team size of two. This limitation was necessary in order to exclude major influence of group dynamics. However, the basic architecture of our system would enable the chatbot to interact with more than two users. The participants received instructions on how to communicate with the bot. They were also introduced to their Google calendars and explained how to schedule an appointment with respect to other events in their diary. Following the design of the moderation strategies, the bot supported them in finding a suitable time slot.

We recruited 40 participants (45 % female) with an average age of 25.6 ($SD = 8.5$). The participants we invited were partly students, but also external candidates with various professional backgrounds to avoid potential bias in the evaluations. Subsequently, we formed teams of two and distributed them evenly among the study conditions. Participation was rewarded with 10€ and the duration of the experiment was about one hour. The study was carried out as a between-subject design. However, the short time span of a single appointment negotiation allows the assumption that the carry-over effect is negligible. To obtain additional data, all teams randomly experienced two moderation strategies during their experiment. Since we manually added participants to the Telegram chat groups, we were able to ensure that all study conditions had the same number of teams assigned and that the sequence of conditions was balanced. A declaration of consent was obtained from all participants before the study was initiated. In this process, participants were informed about information privacy and further use of their data in the evaluation.

The questionnaire was structured in three parts. The first part was conducted prior to the experiment, the second after the interaction with the first system, and the identical third part after the interaction with the second system. All variables were measured with items from validated scales, such as the standardised questionnaires AttrakDiff [4] and MeCue [7]. Items were rated on a 7-point Likert scale, with 1 being the lowest and 7 the highest score. The first part of the questionnaire inquired user characteristics, personal experience with electronic devices, and chatbots in order to identify and measure possible confounding variables. We assume that participants who have a personal aversion to technical devices will not give an unbiased assessment. Since we obtained

the subjects' technological affinity, it was possible to check for confounding group differences. Two participants from different teams had to be excluded from further analysis since they were identified as statistical outliers. The remaining evaluation was therefore carried out with 38 participants.

After the data had been cleaned, we tested for statistical significance between the study groups. The Shapiro–Wilk test showed that almost all emotions and factors are normal distributed ($p \gg 0.05$). The Levene's test indicated that the homogeneity of variances can be assumed ($p \gg 0.05$). However, the one-way-ANOVA could not identify significant differences between the study groups.

The reliability was calculated using Cronbach's alpha for the AttrakDiff and MeCue scales. It was validated with a Cronbach's alpha of 0.86 for 19 items (AttrakDiff) and 0.80 for 24 items (MeCue). Since the Hedonic quality (HQ) shows the stimulation of the user and how much a user identifies themselves with the chatbots, we found differences in the ratings between the study conditions: *Baseline* strategy ($\mu = 4.36, \sigma = 0.7$), *Notification* strategy ($\mu = 4.38, \sigma = 0.7$), *Private Negotiation* strategy ($\mu = 4.83, \sigma = 0.54$), and *Group Negotiation* strategy ($\mu = 4.24, \sigma = 0.68$). In general, the MeCue evaluation showed that users rated the bots as useful and user-friendly. For these scales, the statistical values are as follows: *Baseline* strategy ($\mu = 4.1, \sigma = 2.31$), *Notification* strategy ($\mu = 4.45, \sigma = 1.73$), *Private Negotiation* strategy ($\mu = 5.13, \sigma = 1.78$), and *Group Negotiation* strategy ($\mu = 4.31, \sigma = 1.95$). Again, it can be observed that the *Private Negotiation* strategy was rated the best on average. Being contacted by the system in a private chat to reschedule an appointment seems to have been positively perceived, as it is a less intrusive pattern.

Moreover, we had the participants assess the cognitive load to determine how hard it was for them to follow the task. The Cronbach's alpha was calculated to be an acceptable 0.71 for 8 items. However, the cognitive load of all systems was rated neutral with no significant differences between the study groups ($\mu = 3.5, \sigma = 0.79$). There were also some notable trends related to multi-user specific questions. Users felt that the chatbot treated them equally and fairly in all systems ($\mu = 6.1, \sigma = 1.4$). Additionally, the suggestions of the bot were perceived as helpful for the users' decision making ($\mu = 5.3, \sigma = 1.4$). Acceptance and efficiency thus were rated highly independently of the applied moderation strategy.

5 CONCLUSION AND DISCUSSION

This work presented the implementation and evaluation of four different moderation strategies for a multi-user chatbot. The resulting conversational system was able to participate in a multi-turn group interaction and proved easy to use and robust in a subsequent experiment with two users. We conducted a user study with 40 participants grouped into 20 teams and carried out a quantitative assessment of the perceived acceptance, usability and efficiency of the system. The contribution of this paper is the examination of different moderation strategies under realistic study conditions. Although it was not possible to determine significant differences between the respective strategies, the overall rating of the bots indicated a high level of acceptance among users and a low error rate during interaction, which can be considered high efficiency. The results also showed interesting tendencies, as the assessment of the usefulness correlates highly with the overall rating of the bot. Moreover, participants commented that the system engaged constructively and fairly in the negotiation process. Compared to the baseline strategy, the three remaining moderation strategies led to higher usability ratings. However, users felt less pressured if the suggestions of the chatbot were communicated in a private chat rather than in the group chat. This preference should be considered when designing future group chat strategies, especially since this behaviour also made the conversational assistant appear more intelligent.

Besides, the conversation data were stored in a repository in form of an SQLite database. Our study consequently generated a dataset of a goal-oriented group chat moderated by a chatbot with approximately 2100 exchanges between users and system. This set may help in the further development of chatbots. However, the experiments suggest that additional group members would be beneficial to point out statistical significance. We therefore intend to conduct further studies in the near future where we extent the experiments to a team size of more than two users.

ACKNOWLEDGMENTS

This work received funding within the project "RobotKoop: Cooperative Interaction Strategies and Goal Negotiations with Learning Autonomous Robots" by the German Federal Ministry of Education and Research (BMBF).

REFERENCES

[1] Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. (2020).

[2] Massimiliano Dibitonto, Katarzyna Leszczynska, Federica Tazzi, and Carlo M. Medaglia. 2018. Chatbot in a Campus Environment: Design of LiSA, a Virtual Assistant to Help Students in Their University Life. In *Human-Computer Interaction. Interaction Technologies*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 103–116.

[3] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines* 30, 4 (2020), 681–694.

[4] Marc Hassenzahl, Michael Burmester, and Franz Koller. 2003. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In *Mensch & computer 2003*. Springer, 187–196.

[5] Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. 2020. Bot in the bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[6] Michael McTear. 2020. Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots. *Synthesis Lectures on Human Language Technologies* 13, 3 (2020), 1–251.

[7] Michael Minge, Manfred Thüring, Ingmar Wagner, and Carina V Kuhr. 2017. The meCUE questionnaire: a modular tool for measuring user experience. In *Advances in Ergonomics Modeling, Usability & Special Populations*. Springer, 115–128.

[8] L. Rabe. 2021. Number of digital voice assistants in use worldwide from 2019 to 2024. https://de.statista.com/statistik/daten/studie/1073385/umfrage/anzahl-der-nutzer/von-messenger-apps-weltweit/. https://de.statista.com/statistik/daten/studie/1073385/umfrage/anzahl-der-nutzer-von-messenger-apps-weltweit/

[9] Michael Seufert, Tobias Hoßfeld, Anika Schwind, Valentin Burger, and Phuoc Tran-Gia. 2016. Group-based communication in WhatsApp. In *2016 IFIP networking conference (IFIP networking) and workshops*. IEEE, 536–541.

[10] Saadeh Z. Sweidan, Sarah S. Abu Laban, Njood A. Alnaimat, and Khalid A. Darabkh. 2021. SIAAA-C: A student interactive assistant android application with chatbot during COVID-19 pandemic. *Computer Applications in Engineering Education* 29, 6 (2021), 1718–1742.

[11] Carlos Toxtli, Andrés Monroy-Hernández, and Justin Cranshaw. 2018. Understanding chatbot-mediated task management. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–6.

[12] Vladimir Vlasov, Johannes E. M. Mosig, and Alan Nichol. 2020. Dialogue Transformers.