Response Quality in Human-Chatbot Collaborative Systems

Jiepu Jiang
Department of Computer Science,
Virginia Polytechnic Institute and State University
jiepu@vt.edu

Naman Ahuja
Department of Computer Science,
Virginia Polytechnic Institute and State University
namanahuja@vt.edu

ABSTRACT

We report the results of a crowdsourcing user study for evaluating the effectiveness of human-chatbot collaborative conversation systems, which aim to extend the ability of a human user to answer another person's requests in a conversation using a chatbot. We examine the quality of responses from two collaborative systems and compare them with human-only and chatbot-only settings. Our two systems both allow users to formulate responses based on a chatbot's top-ranked results as suggestions. But they encourage the synthesis of human and AI outputs to a different extent. Experimental results show that both systems significantly improved the informativeness of messages and reduced user effort compared with a human-only baseline while sacrificing the fluency and humanlikeness of the responses. Compared with a chatbot-only baseline, the collaborative systems provided comparably informative but more fluent and human-like messages.

KEYWORDS

Conversational system; chatbot; human-AI collaboration.

ACM Reference Format:
Jiepu Jiang and Naman Ahuja. 2020. Response Quality in Human-Chatbot Collaborative Systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20), July 25–30, 2020, Virtual Event, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3397271.3401234

1 INTRODUCTION

Artificial intelligence (AI) has made remarkable progress in the past decade. For example, in medical imaging and diagnosis tasks, AI can perform as good as trained human experts in an experimental setting [4]. Our society is entering an era where AI and humans will increasingly work together and collaborate. Here we study a representative human-AI collaboration paradigm—both humans and AI can perform the same task with their advantages and limitations, and their collaboration may complement each other.

We examine a particular IR and NLP application area for human-AI collaboration—human-chatbot collaborative conversational systems. Such systems allow users to synthesize their knowledge and chatbot outputs to reply to other people in online text-based conversations. This application problem has many potential commercial and personal use scenarios. For example, e-commerce companies can use such systems to assist a large volume of online customers. An individual user can also benefit from the help of a chatbot to reduce the effort of email and instant message communications.

Our collaborative conversational systems provide users with top-ranked chatbot responses as suggestions. Users may borrow ideas from AI chatbot outputs while formulating their reply messages. This collaboration design is illuminated by query suggestion and auto-completion in web search engines. Previous studies have found that both techniques can help search engine users by reducing the effort to input a query [6] and offering ideas for a new search [3]. We expect chatbot response suggestions to assist users similarly. For example, a user may make a few edits to a chatbot suggestion as a quick reply to save effort. Chatbot outputs may also provide users with relevant information to write their responses when the conversation requires substantial knowledge.

We have designed two collaborative conversational systems and evaluated them using a crowdsourcing user study. We compared the two collaborative systems with human-only and chatbot-only settings in terms of both efficiency and response quality. The rest of the article reports our designs, experiments, and findings.

2 EXPERIMENT

2.1 Collaborative Conversational Systems

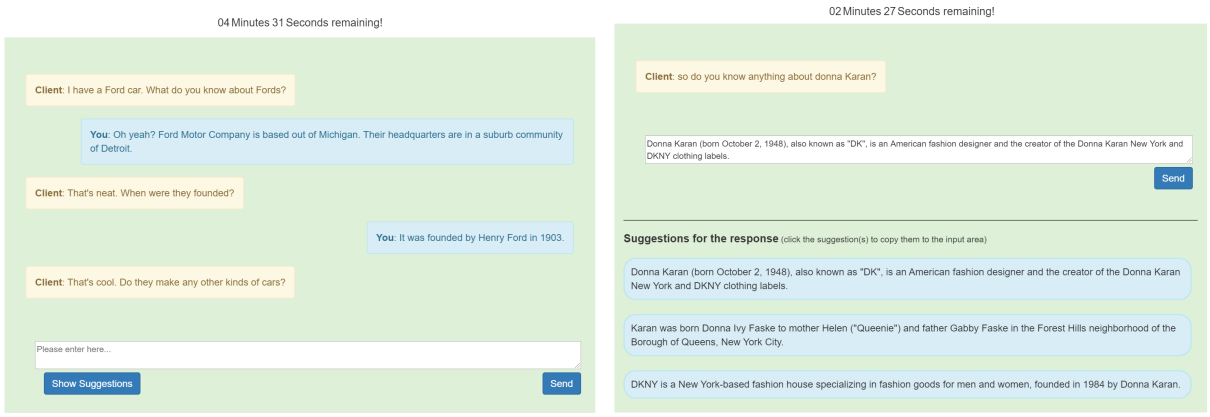
Both our two collaborative conversational systems (C1 and C2) provide users with chatbot response suggestions. But they differ significantly by the extent to which they encourage users to write responses on their own or based on chatbot results.

Figure 1 (left) shows a screenshot of C1. The system provides a "Show Suggestions" button below the input box for entering a reply message. C1 does not show chatbot suggestions until users click on the "Show Suggestions" button. Clicking on a system suggestion will append its content to the current text of the input box. Users can make further edits to the auto-copied content and select other chat outputs. C2 has the same functionality as C1, but C2 presets the input box to the top-ranked chatbot response. Figure 1 (right) shows a screenshot of C2 after clicking "Show Suggestions".

Although C1 and C2 have almost the same functionality, they differ significantly by the extent to which they encourage users to write responses on their own or based on chatbot results. C1 encourages users to write responses by themselves—this is because C1 hides chatbot suggestions by default until users intentionally click on the button. In contrast, C2 encourages users to edit chatbot suggestions to formulate their responses because it presets the text field using the top-ranked chatbot result. C2 "forces" users to examine chatbot suggestions, costing more effort to write a reply message very different from the preset content than an empty text box (as they need to remove lots of the preset texts).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '20, July 25–30, 2020, Virtual Event, China
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-8016-4/20/07…$15.00
https://doi.org/10.1145/3397271.3401234

Figure 1: Screenshots of collaborative conversational system C1 (left) and C2 (right). On the left is a screenshot of C1 before clicking the "Show Suggestions" button. On the right is a screenshot of C2 after clicking the "Show Suggestions" button.



2.2 Experimental Design

We conducted a crowdsourcing user study to evaluate C1 and C2 systems and compare them with human and chatbot-only baselines. Our experiment used a between-subjects design. We assigned each participant to use one of the following seven systems:

- **H** – a baseline human-only system where users are only provided with a text box to input their replies without a chatbot.
- **C1** – three variants of C1 providing one, three, or five chatbot top-ranked responses as suggestions.
- **C2** – three variants of C2 providing one, three, or five chatbot top-ranked responses as suggestions.

We have created a task pool based on the Wizard of Wikipedia dataset [2]. We chose this dataset because the included conversations require some knowledge but not extensive domain expertise (to ensure that crowd workers can handle). The pool consisted of 90 conversations with one, two, or three rounds of utterances (30 for each case). For each task, participants needed to respond to the most recent message. For example, Figure 1 (left) shows an example task with three rounds of utterances. Our experimental system did NOT further reply to the participant's message (which would require either a real human or a human-like chatbot at the back-end). We randomly sampled conversations from the dataset and manually removed those chit-chats or required extensive domain expertise.

We built C1 and C2 based on ParlAI [5], a popular open-source chatbot platform. We adapted Chen et al.'s method [1] to retrieve Wikipedia sentences as chatbot suggestions. We used the Document Retriever [1] to find articles relevant to the current conversation based on up to two previous dialog turns. We extracted the first paragraph of the top three retrieved articles and then chose the top two sentences from the selected paragraphs as suggestions. Figure 1 (right) shows three suggestion responses from Wikipedia. Note that we did not configure the chatbot to retrieve sentences from the Wizard of Wikipedia dataset because we found it would easily push the original response of the conversation to the top. We believe this is unrealistic as it assumes that the corpus always includes a perfect response to a conversation.

We required each participant to finish an experiment session of five minutes (do not count the time spent on instructions and a training task at the beginning). The participants completed conversation tasks randomly sampled from the pool (without replacement) one after another until five minutes. We instructed participants to provide informative responses instead of short and straightforward replies such as "Yes/No" and "I don't know."

We recruited participants from Amazon Mechanical Turk and required them to have a higher than 95% HIT approval rate and at least 1,000 approved HITs. We paid each HIT (a five-minute session) $0.25. We instructed the participants that other human workers would assess their responses, and they needed to finish at least five conversations with informative responses. We instructed them that the top 10% performing HITs (by the number of finished conversations with informative responses) will receive a $0.25 bonus. We determined an informative response as one with an average informativeness rating of at least 2.5 on a 1–4 point scale.

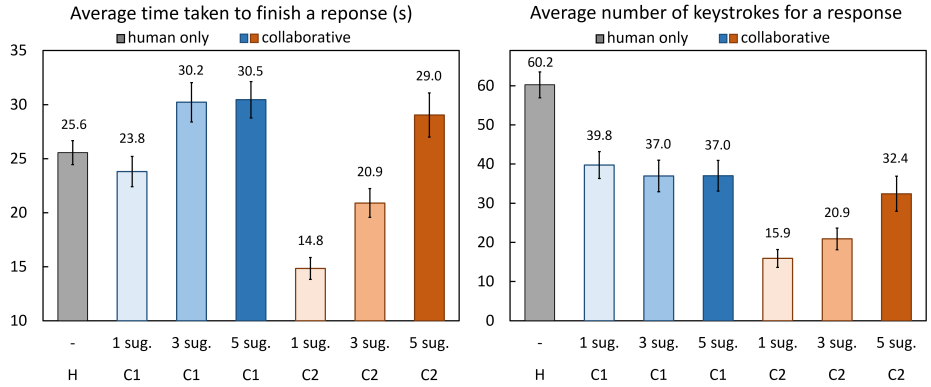
2.3 Response Quality Judgments

We evaluate response quality by *informativeness*, *fluency*, and *humanlikeness*. We collected human judgments on Amazon Mechanical Turk. Each judgment HIT page included the conversation and response to be judged (we have highlighted the response text) and several judgment questions. We paid each judgment HIT $0.04 and required the assessors to have a higher than 95% HIT approval rate and at least 1,000 finished judgments. The questions and options are adapted from previous studies [7, 8]:

- **informativeness** – Does the highlighted response contain sufficient information relevant to the conversation? *Very insufficient* (1), *Slightly insufficient* (2), *Slight sufficient* (3), *Very sufficient* (4).
- **fluency** – How natural is the highlighted response in English? *Very unnatural* (1), *Slightly unnatural* (2), *Slight natural* (3), *Very natural* (4).
- **humanlikeness** – Do you think the highlighted response is provided by a bot or a human? *Definitely a bot* (1), *More likely a bot* (2), *More likely a human* (3), *Definitely a human* (4).

We judged the participants' responses for all the completed conversations in H, C1, and C2 settings. We also used the chatbot for building C1 and C2 to produce responses for the 90 conversation

Figure 2: Time and keystrokes needed for completing a response in human-only (H) and collaborative (C1 and C2) settings.



tasks and collected judgments for the chatbot responses for comparison. We collected three assessors' judgments for each conversation response and used the mean rating as a quality measure.

3 RESULTS

We have collected 140 experiment sessions (20 for each system) and judged the responses in the finished conversations. On average, an experiment session has 10.2 finished conversations. Figure 2 and Figure 3 report the results of the collected data and judgments.

3.1 Response Length, Time, and Keystrokes

Compared with in a human-only system (H), participants in C1 and C2 used less time and fewer keystrokes to provide much longer responses when the systems only showed one suggestion ("1 sug."). This suggests that collaborative systems can improve the efficiency of users to formulate reply messages in a conversation.

Participants' responses in a human-only setting (H) included 52.4 characters on average, compared with 101.1–125.1 characters in C1 ($p < 0.001$ by a t-test) and 140.0–177.5 characters in C2 ($p < 0.001$). Our system has recorded an average of 60.2 keystrokes to finish a response in a human-only setting (H), compared with 37.0–39.8 in C1 ($p < 0.001$) and 15.9–32.4 in C2 ($p < 0.001$). Participants had spent an average of 25.6 seconds to finish a response, compared with 23.8 seconds in C1 (not significant at 0.05 level) and 14.8 in C2 ($p < 0.001$) when the systems showed one suggestion ("1 sug.").

3.2 Response Quality

Crowdsourcing judgments suggest that human-only responses (H) are the least informative but the most fluent and human-like. In contrast, chatbot-only replies are informative but received the lowest fluency and humanlikeness ratings. C1 and C2 collaborative systems reach a balance between the two—their responses are as informative as chatbot-only responses but more fluent and human-like (though not as fluent and human-like as human-only responses).

According to the collected quality judgments, responses in both C1 and C2 settings are significantly more informative than human-only replies (H) and similarly informative to chatbot-only ones. Human-only responses received 2.72 informativeness ratings on average, which is statistically significantly lower than those in C1 (2.81–2.85 depending on the number of suggestions, $p < 0.05$ by a t-test) and C2 settings (2.81–2.89, $p < 0.05$). The human-only responses are also less informative than chatbot-only replies (2.72 vs. 2.78, $p = 0.076$), while C1 and C2 messages received informativeness ratings comparable to chatbot-only responses.

Regarding fluency and humanlikeness, we observed that the human-only setting (H) received significantly higher ratings than both collaborative systems (C1 and C2) and the chatbot-only setting (the differences are all significant at 0.01 level). C1 responses are also significantly more fluent and human-like than chatbot-only replies (all the differences are significant at 0.05 level), while C2 only outperformed chatbot in the humanlikeness of responses.

3.3 Number of Suggestions and C1 vs. C2

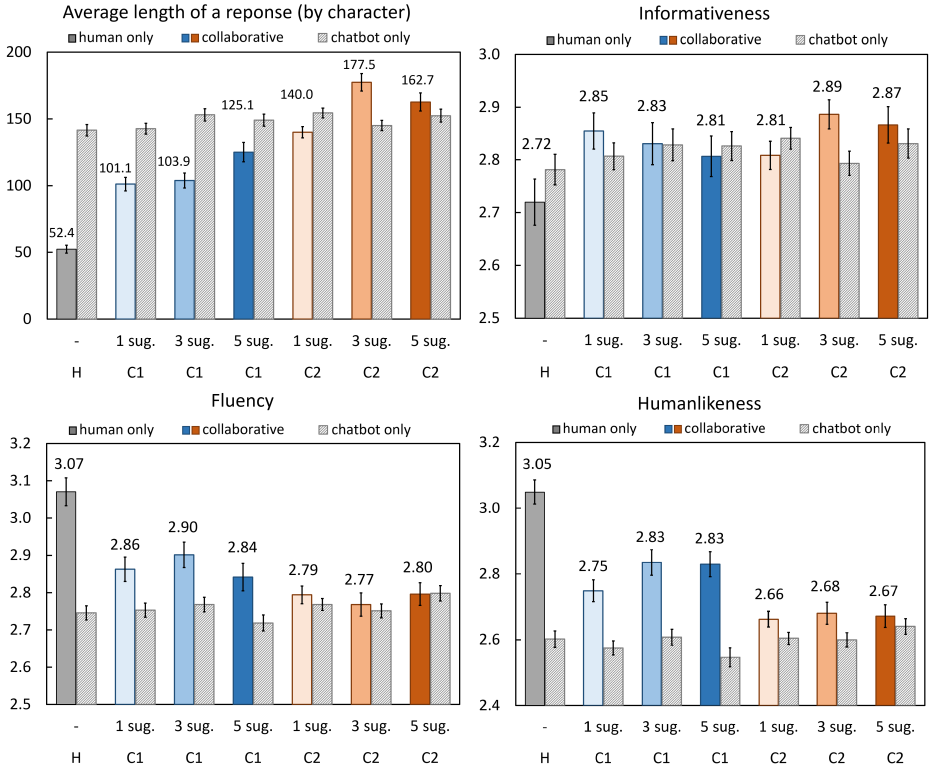
We found that both C1 and C2 systems reach the ideal performance when showing only one chatbot suggestion. Providing more suggestions to users increased the time and keystrokes needed to finish a response but did not consistently improve or hurt response quality. We suspect this is because users needed a substantial amount of effort to examine and synthesize a chatbot suggestion (longer than 100 characters on average in our systems).

For C1 setting, participants spent significantly longer time to finish a response when the system increased the number of suggestions from one (23.8 seconds) to three (30.2 seconds, $p < 0.01$ compared with one suggestion) and five (30.5 seconds, $p < 0.01$ compared with one suggestion). In contrast, the number of keystrokes needed to complete the response did not vary significantly.

For the C2 setting, increasing the number of chatbot suggestions from one to three and five has consistently increased both the time and keystrokes needed to finish a response. The time needed to finish a response increased from 14.8 seconds (one suggestion) to 20.9 (three) and 29.0 (five), where the differences are all statistically significant at 0.01 level. The number of keystrokes needed increased from 15.9 (one suggestion) to 20.9 (three) and 32.4 (five), where the differences are all statistically significant at 0.05 level.

Comparing C1 and C2, users made fewer edits to chatbot suggestions (as indicated from the fewer keystrokes) in C2, and thus the quality of C2's responses are more similar to chatbot-only responses. This is consistent with our expectation, where we preset the text field to encourage users to use chatbot results to a greater extent. This suggests our design has successfully met our purpose.

Figure 3: Response characteristics and quality in human-only (H), collaborative (C1 and C2), and chatbot-only settings.



4 DISCUSSION AND CONCLUSION

We present the first study regarding the efficiency and response quality of human-chatbot collaborative conversation systems to the best of our knowledge. Our work has some important implications:

First, we found that chatbots can help human users improve the informativeness of their responses in conversations requiring knowledge (not chit-chats). We believe this is because well-configured chatbots can provide users with relevant information to help them compose responses. For example, in Figure 1's example tasks, it would be difficult for users to provide informative answers if they do not know much about Ford vehicles and Donna Karan. We expect this finding to generalize to other conversation tasks if the goal is to provide information.

Second, we showed that collaborative systems could improve users' efficiency of composing messages while sacrificing the fluency and humanlikeness of responses. We believe this is because users can make edits to chatbot's suggestions while writing replies, which requires fewer keystrokes but makes the messages less authentic and natural. The low fluency and humanlikeness of responses are also likely affected by our conversational system (which retrieves Wikipedia sentences as suggestions).

Third, our study has demonstrated that we can use simple design choices such as whether to preset the input field to encourage users to synthesize chatbot results by different extents. This difference could also consequently make the results more similar to those in a human-only or chatbot-only setting. It provides an opportunity to balance response quality and efficiency using interface design.

Last, it costs users significant effort (both time and keystrokes, and probably mental effort) to synthesize AI chatbot results with human wisdom. As the results show, providing more chatbot suggestions is mostly negative. This indicates that future system design can benefit from reducing the cost of human-AI synthesis.

REFERENCES

[1] D. Chen, A. Fisch, J. Weston, and A. Bordes. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL '17)*, pages 1870–1879, 2017.

[2] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston. Wizard of Wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations (ICLR '19)*, 2019.

[3] D. Kelly, K. Gyllstrom, and E. W. Bailey. A comparison of query and term suggestion features for interactive searching. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '09)*, pages 371–378, 2009.

[4] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. C. Corrado, A. Darzi, et al. International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788):89–94, 2020.

[5] A. Miller, W. Feng, D. Batra, A. Bordes, A. Fisch, J. Lu, D. Parikh, and J. Weston. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP '17)*, pages 79–84, 2017.

[6] B. Mitra, M. Shokouhi, F. Radlinski, and K. Hofmann. On user interactions with query auto-completion. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval (SIGIR '14)*, pages 1055–1058, 2014.

[7] A. See, S. Roller, D. Kiela, and J. Weston. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of NAACL-HLT*, pages 1702–1723, 2019.

[8] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL '18)*, pages 2204–2213, 2018.