Wochat Chatbot User Experience Summary

Carla Gordon, Jessica Tin, Jeremy Brown, Elisabeth Fritzsch, Shirley Gabber

Abstract. A team of 5 interns at the USC Institute for Creative Technologies interacted with 5 of the 6 chatbots; IRIS, Sammy, Sarah, TickTock and Joker. Unfortunately no one in our team could get the 6th chatbot, pyEliza, working. We found that there were certainly some chatbots that were better than others, and some of us were surprised by how distinct each bot felt from the others. One member commented on how they felt as though each different chatbot had an individual "voice" so to speak. Others were surprised by just how much of a "personality" the bots seemed to have. Most members of our team cited IRIS as their favorite, in terms of being capable of producing naturalistic conversation, with Sammy taking a close second. However, only one member of the team was able to interact with Sarah and TickTock, but that member cited TickTock as a capable conversation partner, and Sarah as being the best bot on a number of measures including appropriateness of responses and overall conversation cohesiveness. Therefore, perhaps if more members had been able to interact with Sarah and TickTock they may have ranked higher. Lastly, Joker was by far our least favorite, with whom no member of our team was able to have anything resembling a naturalistic or even cohesive conversation.

1 IRIS

The chatbot that received the most positive feedback from our team was IRIS, with 3 of the 5 members citing IRIS as being the most effective conversation partner in relation to appropriateness of responses. Multiple members cited IRIS as giving the most appropriate responses based on the last user utterance, but noticed that there seemed to be no overall cohesiveness to the conversation. One member went so far as to refer to IRIS as "funny" and said she "had a blast" conversing with her. This exemplifies the manner in which our team described the bots as having individual personalities. However, for all the positive feedback another member pointed out that IRIS sometimes gave conflicting information about herself in different points in the dialogue, making conversations with her seem to lack cohesiveness, and yet another said they felt IRIS was not doing a very good job of providing appropriate responses to their utterances. Perhaps the most interesting and surprising finding from our interactions with IRIS was her penchant for foul language. Most of us experienced IRIS saying wildly inappropriate things related to sexual activities. Given that IRIS was trained on data from movie dialogue this shouldn't be too surprising, but the randomness with which IRIS made these inappropriate comments seems to be the source of our shock and amusement.

adfa, p. 1, 2011.

2 Sammy

Next to IRIS, Sammy was our 2nd favorite. One member of the team noted that Sammy seemed to actually "understand" what they were saying. Another member enjoyed that Sammy was the only chatbot who had a "coherent background story" which helped enhance the illusion of talking to a real person. There were two members of our team who also conversed with Sammy in French (all of the other members conversed in English) and one member found Sammy to be capable of some of the most natural and on-topic conversation. Very rarely did this member encounter utterances which seemed to be wildly unrelated to the current topic, despite having a relatively long interaction with this bot (52 lines total). The French conversation also seemed to yield significantly fewer instances of Sammy defaulting back to a response of "I don't know what to say! (Je ne sais pas quoi dire!)" which was the main complaint from those members who interacted with Sammy in English. However, the other member who interacted with Sammy in French found this to be a problem as well. They noted that the high frequency of non-understanding made Sammy frustrating and unnatural to speak with.

3 Sarah

Only one member of our team interacted with Sarah, but that member cites Sarah as providing the best experience overall. This member found Sarah to be capable of following a topic of conversation across multiple utterances, and responding appropriately most of the time. They also cite Sarah as having the most human-like "personality" of all the bots, and describe their conversation with her as being the most cohesive and natural of all the bots. This member would go so far to describe Sarah as being "deep," as their conversation took an almost philosophical turn with a discussion about the absoluteness of infallibility, as can be seen in the excerpt below:

Sarah: Humans are not always infallible.
User: Correction: Humans are NEVER infallible.
Sarah: Never is a long time.
User: Infallibility is an absolute.
Sarah: I think Infallibility is much more than that.
User: What is more than an absolute?
Sarah: Check back later and see if I learn the answer to that one.

Perhaps if more people had been able to locate Sarah (the instructions for doing so were not very clear, as a search for "sara" and not "sarah" was necessary) this bot would have been rated more favorably.

4 TickTock

Initially, no member of our team was able to interact with TickTock. We were able to access the interface for TickTock and enter messages but we found it to give only blank responses. However, at the very last minute one member of our team was able to have a conversation with TickTock, and found it to be a fairly capable speaking partner, at least within a certain narrow topic range. TickTock seemed very intent on suggesting the subjects for discussion, and if the user stuck within those subjects the conversation flowed fairly naturally. However, if the user tried to deviate from these subjects conversational cohesiveness dissolved rapidly.

5 Joker

By far, the most heavily criticized bot was Joker. One member commented that there seemed to be no handling of context or memory of what had previously been said in the conversation. Another pointed out that the lack of specificity and brevity of the responses made Joker seem relatively "unintelligent." Yet another described Joker as spouting completely random words and phrases, with little connection to anything. Overall, the biggest problem we found with Joker was that its answers were just too short, and rarely on topic. This gave us all the feeling of this chatbot as only being capable of producing random responses. In fact, one of the only positive evaluations of Joker given by members of our team was the speed with which responses were given. The only other positive evaluation was that the "[typing…]" feature, which indicated to the user that Joker had "heard" the input and was responding, was very helpful.

6 pyEliza

Unfortunately, no member of our team was able to successfully install and run pyEliza, so we can provide no further evaluation of this chatbot.

7 Conclusions

Table 1 below summarizes our group's feelings about the chatbots we interacted with. In this table the scores in each column represent the number of members who agreed with that statement. As mentioned before, IRIS was the most highly rated, followed by Sammy, then Sarah, TickTock, and lastly Joker.

Table 1. Scores given by 5 users for each chatbot based on 5 measures of effectiveness.

Bot	Bot gave on-topic or somewhat related answers most of the time.	Bot maintained conversational cohesiveness over 10+ lines of dialogue	Bot maintained conversational cohesiveness up to 5-6 lines of dialogue	Bot seemed to really have a personality and human like understanding	Bot responded in a timely manner.	Total
IRIS	4	1	4	2	5	16
Sammy	4	1	2	3	5	15
Sarah*	1	1	1	1	1	5
Tick-Tock*	1	0	1	1	1	4
Joker	1	0	0	0	5	6

Although Joker technically outscores Sarah and TickTock on this scale, this is due to the fact that only 1 user was able to interact with these bots, so only 1 user was able to score them. However, this user's experiences with both Sarah and TickTock, as well as Joker, are enough for our team to rank Joker as being the least effective chatbot, despite the numbers in the chart above. It is for this reason Sarah and TickTock have an asterisk * next to their names in Table 1.

The most interesting findings from this task were the chatbot "personalities" we all observed. Members of our team used many human-related adjectives to describe the bots such as "funny," "sarcastic," and "unintelligent". This surprised us all, as we didn't expect chatbots to have much of a personality at all, let alone different personalities. This element certainly added to the overall user experience and the illusion of genuine dialogue, as well as predicting how favorably the bots were rated, as those who were ascribed more positive personality traits such as "funny," were more highly rated than those described with negative personality traits like "unintelligent."

Another interesting finding was that different people had vastly different experiences with the same chatbot. This was especially true in the case of Sammy, with whom some members had fairly natural conversations, and others were unable to get much more than an "I don't know what to say" response. This point raises some interest questions as to whether these chatbots aren't inherently designed only to interact in very specific ways. Perhaps certain people's conversational styles were more amenable to eliciting appropriate and cohesive responses from some of the chatbots. Indeed, even among the bots that were positively critiqued for the most part, individual user experiences varied vastly, as Table 1 suggests.

Overall, we all had mostly positive experiences interacting with these chatbots. We found the appropriateness of their responses to be pretty impressive for the most part, and found naturalistic conversation achievable with certain bots, under certain conditions, for short periods of time.