An Evaluation of Three Chatbots’ Potential as CALL Tools

1. Introduction

Chatbots are agents that interact with users turn by turn using natural language for the conversation purpose (Shawar, & Atwell, 2007). This project employed three chatbots to evaluate their performance as a Computer Assisted Language Learning (CALL) tool. Two English learners were recruited to chat for 40 turns each with three chatbots, namely, Rose, Elbot and Mitsuku. The first subject is a secondary year 2 female student and the second one is a university year 4 male student. Their evaluation of user experience is collected by questionnaire. This paper presents the assessment of annotated utterances, the questionnaire results, and the chatbots’ linguistic qualities. The comparison regarding their performance is presented in section 3 as a summary. Lastly, we outlined the common restraints among the three chatbots and proposed some possible future development.

2. Analysis

1. Conversation assessment

After locating the utterances generated by learners and chatbots into different categories (see Appendix 1 and 2), the ratios of each category are calculated and analyzed.

In terms of subject’s utterance, compared with talking to Rose (23.10%), the average Q rate of these two users goes up when they talk to Mitsuku (28.06%). While the users provided fewer answers to Mitsuku (6.1%) than to Rose (10.57%).

With the highest ratio of general dialogue (3.66%), Mitsuku is also be treated most politely (1.22%).

As for chatbots’ utterance, Mitsuku performed the best with the highest appropriate responses rate (29.27%) and the fewest inappropriate responses (3.66%). Though the other two chatbots behaved comparably, Elbot produced a higher percentage of appropriate responses (25.61%>24.2%) and a lower rate of inappropriate responses (4.88%<5.28%). Rose has much fewer partially inappropriate responses (9.44%<13.42%). Besides, chatbot initiative utterance accounts for 4.88% for Mitsuku, 6.1% for Elbot and 13.72% for Rose, which reflects Rose’s limited ability to stick on ongoing topics.

When comparing two subjects, it can be observed that subject 2 answered chatbots’ questions less frequently (5.34%<10.66%). Subject 2 also produced some hate words (4.64%) while subject 1 did not generate any. The reason may be the disrespect of the more experienced user towards chatbots, resulting in the ignorance to questions and impatience.

B. The questionnaire results

The questionnaire is composed of two questions and the subjects were asked to rate on a scale of 1 (completely agree) to 5 (completely disagree). The results are shown in Table 1.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Rose-1 | Elbot-1 | Mitsuku-1 | Rose-2 | Elbot-2 | Mitsuku-2 |
| This chatbot is fun to talk to. | 4 | 2 | 2 | 5 | 3 | 4 |
| This chatbot helps me learn English. | 4 | 2 | 3 | 3 | 3 | 3 |

Table 1. Questionnaire

It can be inferred that subject 2 generally provided a more negative feedback towards the fun level of chatbots. He gave the same score to all the chatbots in the second question, which demonstrates that subject 2 thought there was less fun in those chats and no difference regarding their helpfulness of learning English.

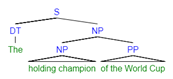
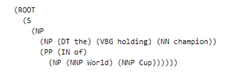
From the aspects of different chatbots, the average score was calculated, leading to the result that Elbot has the best user experience (2.5 in both questions), followed by Mitsuku (3 in both questions). Rose failed in winning the subjects’ favor (4.5 in fun and 3.5 in helpfulness).

C. The chatbots’ linguistic qualities

Besides users’ response, the linguistic qualities of these three chatbots are various in both input comprehension and output generation aspects. Input’s grammatical decomposition performed by these chatbots has been observed and the richness of language use (phrase) is analyzed in the following section.

Chatbots often have to overcome the great difficulty in comprehending human language given its complexity and variability (Hill, Ford, & Farreras, 2015). These problems have been detected to different degrees. In the data collected among the three chatbots, some of the cases demonstrated the unstable comprehension performance. As shown by the response given by Elbot when he responded to Subject 2’s questions of the same type, namely, the subject-auxiliary inversion questions, out of the four responses, Elbot’s answer received an average of 2.25 assessment score. However, two of the provided responses are marked as 3 while one of them was given the lowest score.

Regardless of the unstable comprehension observed, some of the unnatural response can be traced back to the inappropriate grammatical decomposition method the chatbot developers adopted. One of the most noticeable ones is the algorithm used by Mitsuku when she tries to interpret the meaning of phrases. When she attempted to comprehend “the holding champion of World Cup”, instead of decomposing it as a noun phrase consistently, Mitsuku drew an equal line between a noun phrase (champion of World Cup) and a prepositional phrase (of World Cup). Surprisingly, she stuck to the right branch of the tree and decomposed the PP to an NP (World Cup). This misleading method which instructed Mitsuku to always follow the right branch of the grammar tree during parsing changed the root of the phrase (Figure 1). It directly led to the failure in this turn of communication.



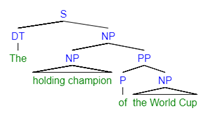


Figure 1: the user input’s parsing and syntax tree.

Delivering the meaning humanlike is another obstacle for chatbots. Struggles have been anticipated while there is some unexpected positive performance such as idiom usage that can elevate users’ daily English usage. The idiom “party killer” generated by Elbot not only delivered the meaning but also avoid a rigid way for the answer, outperforming the other two chatbots in the language use aspect.

Linguistics qualities discussed here from both comprehension and generation perspective confirm the users’ evaluation towards the chatbots. Elbot, which was rated most positively by the subjects, performed some outstandingly natural and helpful response. While there are definitely some flaws existing in the chatbots’ algorithms linguistically such as the syntax tree decomposition.

3. Comparison and Summary

After comparison, Mitsuku may be chosen as the chatbot with the best performance in general and using experience among them, in spite of some obvious grammatical mistakes. It provides more natural responses with proper request repair, encouragement of users’ asking questions and a superior ability to keep to the point instead of digressing from subjects. Elbot, with relatively rich vocabulary, performed satisfactorily in language use and it sometimes surprises users with ingenious responses. Compared with these two chatbots, Rose showed little advantage.

In terms of subjects, the relatively older, more experienced subject with higher English proficiency level tends to show less interest and patience during the conversation because of his higher standards of evaluation. On the other hand, the younger learner found some English knowledge beneficial and helpful during the chat.

Therefore, it is convinced that chatbots have the potential to be an ideal CALL tool for English learners, especially those intermediate-level young ones.

4. Problems and Future Development

Much remains to be done for chatbots being practical CALL tools of large-scale use. Firstly, common sense mistakes such as “current American president is Obama” calls for an update of the database. Secondly, their ability to comprehend input in the context needs to be improved. The incoherent understanding leads to inappropriate responses regarding the above-mentioned topic, especially distant content. A mechanism needs to be developed to extend their memory span. The last but not least, English knowledge provided by chatbots is relatively unorganized, which leads to low learning efficiency. More functions such as hierarchical vocabulary can improve chatbots’ ability to assisted language learning in line with the student’s ability.

Reference

Shawar, B. A., & Atwell, E. (2007, April). Different measurements metrics to evaluate a chatbot system. In *Proceedings of the workshop on bridging the gap: Academic and industrial research in dialog technologies* (pp. 89-96). Association for Computational Linguistics.

Hill, J., Ford, W. R., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, *49*, 245-250.

**Appendix**

**Appendix 1: Evaluation of the user’s utterances**

|  |  |
| --- | --- |
| Code | Meaning |
| A: | Answer to chatbot’s question |
| D: | General dialogue functions, such as greetings, polite expressions |
| H: | Hazing, testing, flaming, such as abusive comments or questions |
| Q: | Question or information request |
| S: | Statement |

**Appendix 2: Evaluation of the chatbot’s utterances**

|  |  |
| --- | --- |
| Code | Meaning |
| 3: | Appropriate response |
| 2: | Partially appropriate response |
| 1: | Inappropriate response |
| RR: | Request repair |
| CI: | Chatbot initiative |