Subjective Annotation and Evaluation of Three Different Chatbots
WOCHAT: Shared Task Report

Naomi Kong-Vega, Shen Mingxin, Wang Mo, Luis Fernando D'Haro

Abstract This paper evaluates the performance of three different chatbots: IRIS, TickTock and Joker, that have been made available to the public online. All three retrieval-based dialogue systems are chat-oriented and designed to engage the users into all types of conversations for as long as possible. They employ different approaches to provide relevant and valid responses, and constantly utilize conversational strategies to further automatically improve its own system through machine learning. The analysis of annotations of more than 2000 responses for the three chatbots allowed us to confirm the robustness, scalability and usability of the systems, as well as to detect a few areas in which response accuracy was lacking, and propose future work to further improve the three systems and annotations scheme.

1 DIALOGUE SYSTEMS

As part of the WOCHAT workshop series, several chatbots are made available to the public online, where users engage in human-machine dialogues. Annotations are then manually generated to see the response accuracy delivered by the various chatbots. The collected dataset and annotations are available to the public and research community for further improvement of chatbot systems. This paper analyzes annotations of a dialogue dataset collected by 3 chatbots in English: IRIS, TickTock and Joker. All dialogs have been collected through volunteers accessing the WOCHAT

Naomi Kong-Vega, Shen Mingxin, Wang Mo
National Junior College, 37 Hillcrest Road, Singapore 288913

Luis Fernando D'Haro
Institute for Infocomm Research, A*STAR
1 Fusionopolis Way #21-01 Connexis South Tower - Singapore 138632
e-mail: `luisdhe@i2r.a-star.edu.sg`

website[1]. Few instructions are given to the participants, i.e. basically to choose the chatbot they want to chat to and keep the dialogue for at least 20 to 30 turns maximum (this is to reduce the complexity and burden of annotating long dialogs), and in case that the chatbot provides a confusing answer, to keep the conversation flow. No indications about the topics or length of the turns are given to allow the participants to be free on exploring the capabilities of each chatbot allowing also the researchers to improve their systems.

1.1 IRIS

IRIS [1] a user-initiative dialogue system that is conversant in a large variety of topics. It is a example-based chatbot which responses are selected from a database consisting of movie quotes, but with the capability of learning new concepts from users. Candidate responses are selected by computing a vector space model representation of both the latest user input and the dialogue history so that IRIS could provide the most relevant responses in the context. Since IRIS is the chatbot that has been available in the website for longer time, it is the one having more number of files for annotation.

1.2 TickTock

TickTock [3] is a rule-based chatbot capable of conducting free-form conversations in contrast of goal-driven systems. Hence, the main objective of this chatbot is to make each conversation as long as possible. Candidate responses are selected from a database of question-answer pairs from 767 CNN interviews. Though the conversations are mainly user-initiated, when the score of the candidate response is below a threshold value, TickTock pro-actively changes the topic by raising a new open question.

1.3 Joker

Joker [2] is a user-initiative dialogue system which employs example-based dialogue modelling, using a database of indexed dialogue examples built from television drama subtitles. Besides providing the participant the most appropriated responses, it also presents social communication skills including humour.

---

[1] http://workshop.colips.org/wochat/main_sharedtask.html

2 EVALUATION OF RESPONSES

For this paper, authors annotated a set of different dialogues from 3 chatbots using the following categories: VALID, ACCEPTABLE and INVALID, according to the guidelines in Table 1, as well as whether there were any offensive or inappropriate swear language for each turn in the dialogue.

VALID	This score is used when the response is semantically and pragmatically valid given the previous recent dialogue history, while allowing few or none grammatical errors. e.g. **USER:** Do you like any other kind of music? **TICKTOCK:** Yes. I like classical music. —-VALID **USER:** What is your favourite song? **TICKTOCK:** Anything I can slowdance to. —-VALID
ACCEPTABLE	This score is given when the response is not necessarily semantically valid but can be acceptable, given the most recent dialogue context, from the pragmatic point of view. This could also mean that the response required slight guessing to make sense OR had severe grammatical errors. e.g. **USER:** I have to get back to work. **JOKER:** I'll go with you. **USER:** Oh that's nice, thank you. **JOKER:** No, we don't live together I mean. —-ACCEPTABLE **USER:** No offense, but I don't want to live with you.
INVALID	Response delivered was completely irrelevant or off-topic, and caused an obvious break in the flow of conversation. e.g. **USER:** Good morning. **IRIS:** Gotta run. Shipment of worms coming in first thing. **USER:** Are you importing worms? **IRIS:** Really? —-INVALID **USER:** You tell me!

Table 1 Criteria for response evaluation categories and examples

Given that there was not full overlap of the dialogues annotated by the authors we could not calculate inter-annotator agreements, but instead all annotations were counted and averaged when yielding the statistics shown in Table 2. In the table we can see that majority of the responses for each of the 3 chatbots were VALID, therefore the response accuracy for all of them was relatively high.

2.1 IRIS

Each dialogue comprised of an average of 31 dialogue turns between the user and IRIS. With the highest rate of VALID responses of 79.7% and the lowest rate of

	IRIS	TickTock	Joker
No. of different responses annotated	1760	738	1060
(%) of VALID responses	79.7	73.5	63.3
(%) of ACCEPTABLE responses	9.1	10.8	14.1
(%) of INVALID responses	11.2	15.7	22.6
(%) of offensive or inappropriate swear language	1.4	0.1	0.5

Table 2 Statistics for response evaluation for the 3 chatbots

INVALID responses of 11.2%, dialogues with IRIS had the best response accuracy among the 3 chatbots evaluated. IRIS was found to respond well and carry on relatively longer conversations in political topics, and general conversation about feelings and emotions as seen in Table 3.

Political Topics	**USER:** yes, you know in trump's last rally **IRIS:** None of the papers covered the rally. Not even The Sun. **USER:** but it's all over the web **IRIS:** yes, it is.
General Conversation	**USER:** you are scaring me **IRIS:** I'm scaring myself, because I think so much of her **USER:** is she pretty? **IRIS:** She is **USER:** this is so romantic **IRIS:** Mmm. **USER:** Do you love her? **IRIS:** [USER's Name]...

Table 3 Examples of VALID responses by IRIS

On the other hand, IRIS was found to be weak at relating to statements made a few turns before, and usually responded well most of the times based on the latest respond it got from the user. This also lead to abrupt changing of topics in cases that it was unable to understand the users' most recent response. IRIS was also unable to retain information (see Table 4). For example, even if the user had stated his name a few turns ago, it would address the user with another name.

Inability to relate to Contextual Statements	**IRIS:** And how, and how, can these death traps be avoided? **USER:** Which death traps? **IRIS:** Which is? **USER:** I have no idea, that is why I am asking you **IRIS:** I have no idea

Table 4 Examples of INVALID responses by IRIS

2.2 TickTock

Each dialogue comprised of an average of 25 dialogue turns between the participant and TickTock. With the rate of VALID responses at 73.5% and the rate of INVALID responses at 15.7%, dialogues with TickTock had the second best response accuracy among the 3 chatbots evaluated.

TickTock, unlike other chatbots, usually initiates conversations with questions that introduce a topic it would like to talk about. This allowed the users to set the context for TickTock to respond appropriately to. TickTock was found to respond well and carry on longer conversations related to video games, music, or films, but especially general conversation about the users' life. This is because unlike the other chatbots, TickTock responds with vague and open questions, allowing the users to lead the conversations (see Table 5).

Initiating conversations with topics	**TickTock:** Can you tell me something about politics? **USER:** Trump will make America great again. **TickTock:** Politicians and diapers have one thing in common. They should both be changed regularly, and for the same reason. **USER:** That's a good one.
Vague yet relevant responses	**TickTock:** What's your idea? **TickTock:** I don't quite understand. Any thoughts?

Table 5 Examples of VALID responses by TickTock

TickTock was found to be generally weaker at answering open questions asked by the user. In a few cases too, though with similar strategy of using vague responses, it produced completely irrelevant responses (see some examples in Table 6).

Vague and irrelevant responses	**USER:** I don't ski **TickTock:** I certainly try. What do you think? **USER:** How old are you? **TickTock:** Uh-huh, I do. What do you think?

Table 6 Examples of INVALID responses by TickTock

Like IRIS, TickTock was unable to keep up with the context of the dialogue, and could only respond based on the latest response of the user. It was also unable to retain information, as users had to repeat what the topic of discussion was in several occasions.

2.3 Joker

Each dialogue comprised of an average of 42 dialogue turns between the participant and Joker. With the rate of VALID responses at 63.3% and the rate of INVALID responses at 22.6%, dialogues with Joker had the lowest response accuracy among the 3 chatbots evaluated.

Joker was found to provide accurate responses to short and direct questions, and responded with brief comments that were still moderately relevant so the dialogues made sense, and could continue (see some examples in Table 7).

Responses to short and direct questions	**USER:** how are you? **Joker:** good **USER:** What did you do today?. **Joker:** nothing
Moderately relevant brief comments	**USER:** That's great news! I'm proud of you **Joker:** mm-hmm

Table 7 Examples of VALID responses by Joker

Unfortunately, Joker mostly responded with very short statements, and did not show initiate in the conversation. Thus, even though the dialogues kept going, there was very frequent changes in topics, and it waited for the participant to feed it information to respond to, repeating other times what was said or providing extremely short statements. Furthermore, there were many cases where Joker's responses were irrelevant and random, making dialogues confusing and short-lived (see examples in Table 8).

Very short statements that caused frequent change of conversation topics	**USER:** I am feeling sick today **Joker:** well that's **USER:** how are you feeling today? **Joker:** okay **USER:** what's the weather like? **Joker:** good
Completely irrelevant and random responses	**USER:** so what's up today? **Joker:** they lived together **USER:** who lived together? **Joker:** where's Gordon and has he changed? **USER:** I have no idea who Gordon is

Table 8 Examples of extremely short & INVALID responses by Joker

Like IRIS and TickTock, Joker was unable to retain and relate to previously mentioned information. Its biggest weakness was that it only gave short responses making the users to essentially come up with conversation starters very often.

3 CONCLUSION AND FUTURE WORK

In this paper, we reported the results of annotating and subjectively evaluating the responses of three different chatbots - IRIS, TickTock and Joker. Our expectation is that by providing this analysis, we can highlight the major changes that must be done to these chatbots, but also to provide a set of important aspects that any non-oriented dialogue chatbot should have to keep engaged the users.

All chatbots were fairly proficient in giving valid responses, with all of them only having a maximum of 25% invalid responses. Unfortunately, all chatbots were found to be weak in retaining and relating to information from a few statements back, resulting in frequent changing of topics, and thus, no engaging conversations. A summary of the performance of the chatbots is given in Table 9.

System	**IRIS**	**TickTock**	**Joker**
Processing of context information	Yes	No	No
Conversation length / dialogue turns	31	25	42
Retain of information	Only seems to retain information from the last user's turn	Only seems to retain information from the last user's turn	Only seems to retain information from the last user's turn
Strengths	Respond well and sustain relatively longer conversations related to politics and emotions.	Initiating conversations; good responses to topics on video games, music, films and user's life.	Responding to short and direct questions well; providing short comments to sustain the conversations.
Weaknesses	Could not retain information. Could not respond in context.	Could not retain information. Most responses to open questions were vague and irrelevant.	Could not retain information. Responses are very short and do not initiate conversations.

Table 9 Summary of performance of chatbots

As future work, we plan to help developers to go beyond response appropriateness (i.e. a higher rate of VALID responses) by improving their chatbots' consistency in adhering to the conversation's general context, and also increasing the engagement level of the responses. As such, we plan to increase the level of annotation by providing alternative answers for the INVALID and ACCEPTABLE cases.

Besides, we are planning to provide additional annotations like the exact turn where specific information was provided. This way, if the user mentioned his name in a given turn, the annotation will point to that turn allowing training the chatbot to pay attention to the correct context in the dialogue. Finally, we expect more annotations and annotators could be used to calculate inter-annotator agreements, automatically evaluate the responses, detect different kind of non-appropriated interactions, among others.

4 ACKNOWLEDGEMENTS

We would like to thank Dr. Rafael E. Banchs from Institute for Infocomm Research (I2R) for providing us the chatbots data used in this project, and Mr. Kester Wong from National Junior College for providing this valuable opportunity to learn about Infocomm Technology that as students would otherwise not have been able to such depths.

References

1. Banchs, R.E., Li, H.: Iris: a chat-oriented dialogue system based on the vector space model. In: Proceedings of the ACL 2012 System Demonstrations, pp. 37–42. Association for Computational Linguistics (2012)
2. Duplessis, D., Letard, V., Ligozat, A.L., Rosset, S.: Joker chatterbot re-wochat 2016-shared task chatbot description report. In: RE-WOCHAT: Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents-Development and Evaluation Workshop Programme (May 28 th, 2016), p. 45
3. Yu, Z., Papangelis, A., Rudnicky, A.: Ticktock: A non-goal-oriented multimodal dialog system with engagement awareness. In: Proceedings of the AAAI Spring Symposium (2015)