

# A Subjective Evaluation of Chatbot Engines

Amanda Cercas Curry and Verena Rieser

Interaction Lab, Heriot-Watt University, Edinburgh, UK  
{ac293, v.t.rieser}@hw.ac.uk

**Abstract.** We participated in two roles in this shared task: Data generator and data annotator. As data generators we were interacting with a number of different chatbots. We find that data-driven systems can produce inappropriate and out-of-character responses, especially when based on TV drama scripts, whereas hand-crafted bots are more able to respond in a context-appropriate manner, but are limited in their scope. We also provide some feedback on using the annotation scheme.

**Keywords:** chatbots, dialogue, subjective evaluation, data annotation

## 1 Task 1: Data Generation

The first task involved chatting to the available chatbots in order to generate human-bot dialogues. In the following, we report on our observations using different chatbots.

### 1.1 Joker

The Joker chatbot is an example-based system that uses a database of indexed dialogue examples automatically built from a corpus of TV drama subtitles to manage social open-domain dialogue.

Joker uses a lot of responses from TV drama scripts, often referring to the user as “Ms Tascioni”, “Mrs Florrick” and “Diane”, all characters from “The Good Wife”. While this can be humorous at times, see example in Table 1, particularly if the user is familiar with the TV drama in question. However, the responses seem to be a bit random at times and do not make sense within the context of the conversation, see Table 2. Joker also tends to give blunt responses to user questions and uses swear language, so it can come across as being rude and uninterested in the conversation. This can make it hard to maintain a conversation with the chatbot. On a scale from 1 to 5 (with 5 indicating “most enjoyable conversation”), we would score Joker 1.

### 1.2 IRIS

IRIS (Informal Response Interactive System) is a chat-oriented dialogue system based on the vector space model framework. Her character is described as being

“like a nice teenager that likes to chat to people but sometimes loses her temper and gets a little naughty”.

Similar to Joker, IRIS has been trained on film data, however IRIS is able to ask the user questions, prompting further user input and making the conversation flow more easily. Also like Joker, IRIS often makes inappropriate remarks, see examples in Table 3, 4, suggesting film and TV data is not necessarily the best resource of impersonating a “nice teenager”. We score IRIS 2/5.

### 1.3 Sammy

Sammy is a chatbot based on the public “small-talk” domain available at api.ai. Users can choose to interact with her in English, French or Italian. Sammy introduces herself as an assistant chatbot and often offers to help the user with work.

Sammy’s responses are more coherent than the two previous chatbots. They are generally grammatically correct and complete sentences. However dialogue with Sammy is very limited. Although she is presented as an assistant and constantly asks for work to do, she does not have many everyday functions that she can perform, like telling the time, see example in Table 6. Also, Sammy is purely task-driven and reluctant to engage in social interaction. Out-of-domain questions are repeatedly met with “I do not know what to say!”, which makes Sammy sound less natural, see Table 5. We rate Sammy 2/5.

### 1.4 PyEliza

PyEliza is a Python-based stand-alone version of the famous Eliza chatbot created by Weizenbaum in 1966, with the intention of simulating a Rogerian psychotherapist. Eliza follows simple pattern-matching techniques and follows a determined script.

PyEliza works by asking the user follow-up questions in a non-directional way. This makes the user keep talking and keeps the conversation flowing, since the questions seem less out of place than other responses described before, see example in Table 7. However, since pyEliza is following a script, she is not able to answer any user questions and tries to draw deep understandings of even the most basic interactions which can be frustrating, see Table 8. PyEliza scores 3/5.

### 1.5 Sara

Sara is a version of Alice bot, developed by Dr. Wallace in 1995. It is based on the AIML framework, which works through pattern matching.

Sara is able to ask the user questions that are relevant to what the user said earlier, so the interaction flows and Sara seems genuinely interested in chatting to the user, even if some interactions still seem unnatural. In Table 9, the user tells Sara that they live in London and Sara responds by asking where she can get “spotted dick”, a British pudding. While this demonstrates a broader knowledge of British culture, it causes an abrupt change of context. We rate Sara 4/5, which makes her our favourite bot to talk to.

## 2 Task 2: Data Annotation

The second task consisted of annotating dialogues. We annotated a total of 16 dialogues and 492 data instances following the annotation scheme provided by the organisers. Data instances were to be marked according to their validity, polarity, whether they were offensive and whether they contained swear language.

The data was annotated using the annotation scheme provided, which presented some challenges, mainly in the two areas outlined below.

**Validity of a response:** A response could either be valid, acceptable or invalid, taking into account its context. Valid responses were semantically and pragmatically valid. Acceptable responses were pragmatically valid even if they were not semantically valid. Examples were provided for each case, however it was unclear what to do about responses which were grammatically incorrect but could be interpreted to be a valid or acceptable response.

**Polarity of a response:** The annotation scheme also invited annotators to use optional tags for the polarity of a response. Because these were optional it was unclear how polarised a response should be in order to be marked. No examples were provided of negative or positive polarities so it was completely left up to the annotator to decide.

## References

1. Guillaume Dubuisson Duplessis, Vincent Letard, Anne-Laure Ligozat, and Sophie Rosset *Joker Chatterbot*. Re-WoChat 2016 - Shared Task Chatbot Description Report, 2016.
2. Rafael E. Banchs, and Haizhou Li *IRIS (Informal Response Interactive System)*. Re-WoChat 2016 - Shared Task Chatbot Description Report, 2016.
3. Luis Fernando D'Haro *PyEliza: A Python-based implementation of the famous computer therapist*. Re-WoChat 2016 - Shared Task Chatbot Description Report, 2016.
4. Bayan AbuShawar *Sarah Chatbot*. Re-WoChat 2016 - Shared Task Chatbot Description Report, 2016.

Joker: "You see the puppy"  
User: "What puppy?"  
Joker: "You"  
User: "Cute"  
Joker: "Thanks offer a plea bargain"

**Table 1.** Joker: Contextually inappropriate response.

IRIS: "How's the tea?"  
User: "It's lovely."  
IRIS: "What's the matter?"  
User: "Nothing, I'm just a bit tired."  
IRIS: "Just a little bit?"

**Table 3.** IRIS: Prompting user input.

User: "Do you like to watch films?"  
Sammy: "I don't have an opinion on that. I like to stay neutral on certain topics."  
User: "Do you like music?"  
Sammy: "I'd prefer to keep quiet on that subject. A good assistant shouldn't be too opinionated."

**Table 5.** Sammy: Example of small talk.

User: "How are you?"  
pyEliza: "What is it you're really asking?"  
User: "How are you feeling?"  
pyEliza: "Perhaps you can answer your own question."

**Table 7.** pyEliza: Open follow-up questions.

User: "Where do you live?"  
Sara: "I am living in unknown. Where are you?"  
User: "I am in London"  
Sara: "Where can I find spotted dick?"

**Table 9.** Sara: Example of abrupt context shift.

User: "Hello"  
Joker: "Hello"  
User: "How are you?"  
Joker: "I have to get back to work"  
User: "What is your job?"  
Joker: "My tampon"

**Table 2.** Joker: Humorous response.

User: "Hi Iris"  
IRIS: "Hi"  
User: "What are you doing?"  
IRIS: "What is this [SWEAR LANGUAGE]?"

**Table 4.** IRIS: Out-of-character response

Sammy: "I'm here to make your life easier. What can I help you with today?"  
User: "What is the time?"  
Sammy: "I do not know what to say!"  
User: "What is the definition of table?"  
Sammy: "I do not know what to say!"

**Table 6.** Sammy: Task-assistant.

pyEliza: "Nice to meet you [User]. How is it going?"  
User: "Good!"  
pyEliza: "Why do you say that good?"  
User: "The weather is very nice."  
pyEliza: "Ok, but can you elaborate a bit?"  
User: "It has been raining for a few weeks."  
pyEliza: "Ok, but can you elaborate a bit?"

**Table 8.** pyEliza: Repeated replies.

User: "How many people are you chatting to?"  
Sara: "I've chatted to half a million people. I can chat to about millions people at once"  
User: "That is amazing"  
Sara: "It seems pretty amazing to me too"

**Table 10.** Sara: Contextually appropriate replies.