Chatbot System for Open Ended Topics Based on Multiple Response Generation Methods

Anna Maria Hadjiev
Graduate School of Information Science and Technology, Hokkaido University
Sapporo, Japan
hadjiev@wisc.edu

Kenji Araki
Faculty of Information Science and Technology, Hokkaido University
Sapporo, Japan
araki@ist.hokudai.ac.jp

Abstract

This paper presents a chatbot intended for open-ended conversations in which the user is free to chat about any topic. The chatbot contains four separate methods of generating responses, three rule based and one based on deep learning. Based on the characteristics of the user input, the system chooses the best of the four methods to produce a response. Evaluations on the system's ability to have satisfying conversations with the user were conducted by five participants. The results support that, in the context of free conversations, a chatbot benefits in its ability to hold satisfying conversations by having multiple response generation methods built in.

Key Words
Chatbot, rule base, deep learning, genetic algorithm, inductive learning

1. Introduction

Nowadays, chatbots are commonly seen for tasks whose interactions are constrained within set themes, like scheduling meetings and finding a restaurant. However, it is rare to see chatbots used for more open ended tasks such as casual conversation. The reason for this is because chatbots struggle when they are used for tasks where the system has to be able to handle a wider variety of possible inputs. One key technique chatbots use for response generation is following rules predetermined by the developers in the form of "if the user inputs x, then the system outputs y," in which x and y are both specific, predefined text. This technique fails when the user inputs something that cannot be matched with a rule, and with an unpredictable task like casual conversation, it is impossible to prepare enough rules to cover every input [1].

Unlike the rule based method, chatbots which rely on deep learning techniques to generate responses are able to provide a response for any input. However, compared to outputs of rule based chatbots, responses generated by deep learning techniques are less predictable and carry a tendency to produce nonsensical or inappropriate responses [2]. For the task of creating a chatbot which can handle a wide variety of inputs, it would be beneficial to incorporate the flexibility of deep learning methods as well as the accuracy of rule based methods.

In this paper, we develop a chatbot system which uses three separate rule based methods as well as a deep learning method. For each input, one of the four methods is selected to generate a response. We define the best response to be one which can best help provide a satisfying conversation, measured by an average score measured across a few factors.

2. Outline of System

As mentioned in the previous section, we implement four different methods of generating responses, three of which are based upon rules and one which is based on deep learning. Based on the user input, the system determines which method would produce the optimal response and generates a response based on that decision (see Figure 1). The four different methods are simple rule base, GA-IL [3], Deep Learning [2] and ELIZA [5].

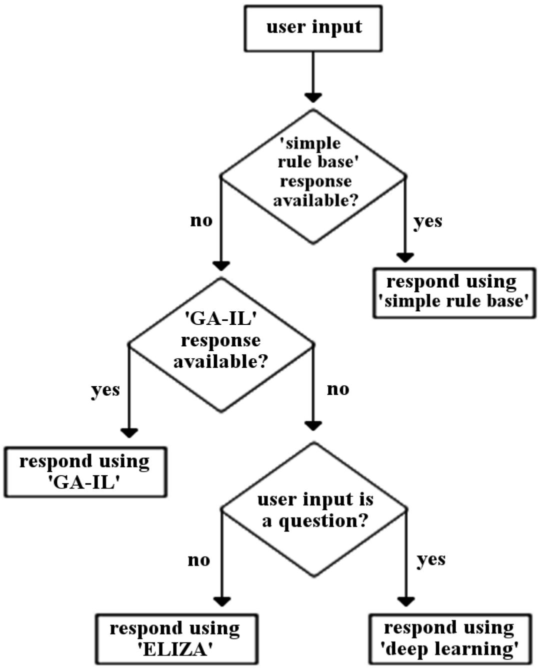

user input
'simple rule base' response available?
no
yes
respond using 'simple rule base'
'GA-IL' response available?
yes
no
respond using 'GA-IL'
user input is a question?
no
yes
respond using 'ELIZA'
respond using 'deep learning'

Figure 1. Flowchart showing the system's process of generating a response to an input

2.1 Simple rule base method

For the simple rule base method of response generation, we rely on rules that we predetermined and hard coded into the system. If the user input matches any of the rules we put in, then the matched rule is followed to generate a response.

Table 1: Examples of rules used to respond

User input	Response
"How old are you" or "What is your age"	"I am a couple months old."
"How are you?"	"I'm doing great, thank you."

Of the four main methods implemented in this system, the rule based method is expected to produce the best results and thus, whenever possible, this method is used over the three others. Unfortunately, this method is not flexible and most input cannot be matched to any of the written rules - when an input cannot be handled by this method, we move onto the next potential method.

2.2 Inductive Learning with Genetic Algorithm (GA-IL)

For this method, the system also relies on pre existing rules to determine the output - however, a key difference is that these rules are learned as the system observes interactions between itself and the user. The algorithm used to generate these rules is based on a method of the same name, GA-IL (Araki et al., 2006) [3], which has produced positive results when applied in casual conversation.

The algorithm works by finding similar patterns between pairs of an input sentence and its reply, and generating new rules to use based on this information. An example of this process is given below in Figure 2.

1. Pairs of a sentence and its response
   a. 'Do you like tennis?' : 'Yes I like it.'
   b. 'Do you like cats?' : 'Yes I like them.'
2. Rules generated from 1:
   Do you like @1 : Yes I like @1
   tennis : it , cats : them

Figure 2. Example of rule acquisition using GA-IL

With the GA-IL method, the system begins with no rules - thus, it requires to be run alongside a different response generation method until it populates enough rules to trigger. Because it starts with zero rules, there would not be enough rules for GA-IL to be used if we do not pretrain it to have rules from the beginning. Thus, we used a dataset provided by a chatbot challenge called DBDC3 [4] which contains conversational data between participants and chatbots, annotated for correctness. Going into the experiment, the GA-IL system contained 289 rules generated using the DBDC3 dataset.

2.3 Deep Learning

The third method we incorporated into our system is deep learning. For this method, we used RNNs (recurrent neural networks), specifically a dual LSTM neural network to create a generative-based model [2]. This means that the system generates original text unlike rule based methods which are more focused on choosing from a set of predetermined text. We chose this implementation because this method has been shown to produce models which are able to learn semantically and syntactically meaningful representations of sentences [2].

For the dataset we used the Cornell Movie Dialogs Corpus, which contains 220,579 exchanges taken between 10,292 pairs of movie characters, and our model was developed using Tensorflow.

```
RNN cell type: LSTM
RNN size: 25
Bi-directional encoder: true
Encoder layers: 2
Encoder embedding size: 256
Decoder layers: 2
Decoder embedding size: 256
Epochs: 500
Batch: 128
Learning rate: 2.0
```

Figure 3. Model and training parameters

This method is less predictable than the previous rule based methods, however it is able to handle any input by the user. Thus, we use this method whenever the user input does not match a rule and is a question. In the case where the user input does not match a rule and is not a question, the next method ELIZA is used because it performs extremely well against non-inquiring inputs and can handle any input.

2.4 ELIZA

One rule-based chatbot implementation which is able to successfully handle almost any input is a well known program created by Joseph Weizenbaum at the MIT Artificial Intelligence Library called ELIZA [5]. This program takes the user input and responds by directing the user to expand upon what they said, in a sense creating an illusion of a conversation.

User: How are you?

ELIZA: Why do you ask?

User: Because I am interested

ELIZA: Is that the real reason?

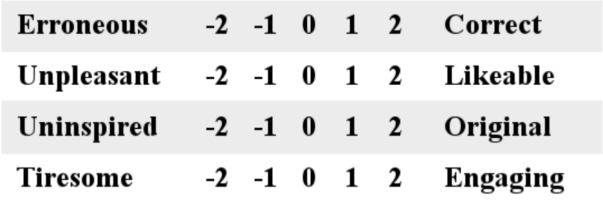
Figure 4. Example of chat via ELIZA

Although this method is highly flexible and can handle most inputs, because ELIZA is unable to provide its own thoughts and only the user meaningfully contributes to the conversation, chatting with ELIZA does not make for a very satisfying conversation. Especially in the case that the user asks a question, ELIZA is unable to give a satisfying response. Thus, we put this method behind rule based and GA-IL.

3 Experiments for evaluation

For our experiment, we recruited five participants: four males and one female, all in their early twenties who are currently working in the I.T. industry and whose native language is English. We had the participants chat for 30 turns with a system consisting of only the deep learning method, a system consisting of GA-IL and ELIZA (ELIZA set to run only when GA-IL could not), and finally our complete system. After chatting with each system, the participants were asked to score their experience based on a few different factors intended to measure how satisfying their conversations were: correctness, likeability, originality, and engagingness.

For the evaluations, we presented the four factors in a semantic differential scale. The semantic differential scale, which presents a rating in the form of a description and its antonym on the other side of the scale, is an effective way to have participants give ratings in evaluations.



Erroneous	**-2**	**-1**	**0**	**1**	**2**	**Correct**
Unpleasant	**-2**	**-1**	**0**	**1**	**2**	**Likeable**
Uninspired	**-2**	**-1**	**0**	**1**	**2**	**Original**
Tiresome	**-2**	**-1**	**0**	**1**	**2**	**Engaging**

Figure 5. Semantic differential scales that the participants used for evaluations

4.2 Evaluation results

The participants conversed with out chatbots for 30 turns each, and the results of their evaluations of the chatbots are given below in tables 2, 3 and 4 for a chatbot using only the deep learning method, a chatbot using GA-IL combined with ELIZA and out complete system, respectively.

Table 2. Evaluations on deep learning chatbot

	A	B	C	D	E	avg.
Correct	-1.0	-1.0	-2.0	0.0	-1.0	-1.0
Likeable	0.0	1.0	0.0	2.0	-1.0	0.4
Original	1.0	2.0	1.0	0.0	1.0	1.0
Engaging	-1.0	0.0	1.0	0.0	-1.0	-0.2
avg.	-0.3	0.5	0.0	0.5	-0.5	

Table 3. Evaluations on GA-IL + ELIZA chatbot

	A	B	C	D	E	avg.
Correct	2.0	2.0	1.0	2.0	2.0	1.6
Likeable	-1.0	-2.0	-1.0	-1.0	0.0	-1.0
Original	-2.0	-1.0	-2.0	-1.0	1.0	-1.0
Engaging	2.0	0.0	-1.0	2.0	0.0	0.6
avg.	0.3	-0.3	-0.8	0.5	0.8	

Table 4. Evaluations on chatbot of all 4 methods

	A	B	C	D	E	avg.
Correct	1.0	0.0	1.0	1.0	1.0	0.8
Likeable	0.0	2.0	0.0	1.0	1.0	0.8
Original	1.0	1.0	0.0	2.0	0.0	0.8
Engaging	1.0	1.0	0.0	1.0	0.0	0.6
avg.	0.8	1.0	0.3	1.3	0.5	

Table 5. Average ratings given by participants A, B, C, D and E for each chatbot

	A	B	C	D	E	avg
Deep learning	-0.30	0.50	0.00	0.50	-0.50	0.04
GA-IL + ELIZA	0.30	-0.30	-0.80	0.50	0.80	0.10
All 4 methods	0.80	1.00	0.30	1.30	0.50	**0.78**

5. Conclusion

Looking at the results of our experiment, we were able to show that chatbots intended for satisfying casual conversations benefit from containing multiple response generation methods to choose from depending on the user input. The results were thanks to the methods combining strengths in a way as well as making up for each other's weaknesses, especially in a context like casual conversations which involve unpredictable interactions.

Running a chatbot based on predetermined rules of what exactly to respond given certain inputs makes for the best results - however, in our experiment these rule based methods were unable to be used frequently due to the lack of rules. In the future we hope to address this in ways such as expanding upon the set of rules we hard coded into the system. As for the GA-IL method, we hope to pretrain the system more on conversational data so that more rules are saved. Furthermore, in the future we are open to incorporating additional response generation methods on top of the four we already have.

References

[1] Maali Mnasri, "Recent advances in conversation; NLP: Towards the standardization of Chatbot building," Mar. 21 2019. arXiv:1903.09025v1 [cs.CL]

[2] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," Sep. 3 2014. arXiv: 1406.1078v3 [cs.CL]

[3] Kenji Araki, Michitomo Kuroda, "Generality of spoken dialogue system using SEGA-IL for different languages." *Proceedings of the IASTED International Conference on COMPUTATIONAL INTELLIGENCE*, pp.70-75, San Francisco, CA, U.S.A., 2006.

[4] R. Higashinaka, K. Funakoshi, Y. Tsunomori, T. Takahashi, N. Kaji, DBDC3, GitHub, https://dbd-challenge.github.io/dbdc3/datasets.html

[5] J. Weizenbaum, ELIZA - A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9(1)*, pp.295-300, 1966.