

AN EVALUATION OF A MULTILINGUAL CHATBOT

Mahesh Vanjani, Texas Southern University, mahesh.vanjani@tsu.edu
Jamison Posey, University of Mississippi, jposey@bus.olemiss.edu
Milam Aiken, University of Mississippi, maiken@bus.olemiss.edu

ABSTRACT

Chatbots are computer programs that seek to emulate human dialogue, and thus, provide a more natural user interface. Although some support other languages, most converse only in English, and only a few can communicate in multiple languages. Here, we describe a new computer application that links the Rose English-speaking chatbot with Google Translate, thus providing conversational capability in 103 languages. To evaluate this new system, a group of undergraduate students reviewed responses from either it or a human, without knowing which was which in a modified version of the Turing Test. Input text to the application was automatically translated from German to English for the Web-based Rose chatbot to process. Results show that the students often were able to ascertain which responses came from the chatbot, showing that it has not quite reached human quality. Regardless, people can communicate with a chatbot effectively in a different language with this hybrid application.

Keywords: Chatbot, Multilingual, Communication, Conversational Agent

INTRODUCTION

People are using chatbots more frequently as a means of communicating with computers, and one study predicts that 80% of all businesses will use these programs by 2020 (Elezaj, 2018). About 69% of consumers prefer these systems for quick communication, and 40% do not care if an interaction is with a human or chatbot as long as they receive the assistance they require. Since talking and texting with mobile apps is so natural, many consider these conversational agents as a preferable interface to applications (Lommel, 2018; Oppenheimer, 2016).

Not only do the systems allow a more intuitive interface with applications to retrieve information or perform actions, but also, users can communicate with chatbots for entertainment, therapy, or even foreign language practice (De Gasperis and Florio, 2012; Fryer, et al., 2017; Goda, et al., 2010; Kerry, et al., 2009; Kreisa, 2018). Now, developers are integrating multilingual translation technology with these programs to enable users to communication in their native languages (Nedrya, 2016; Tiwari, 2018).

This paper describes a new hybrid chatbot that provides communication with a chatbot in 103 languages in 10,506 combinations. First, we provide a background on the technology and then evaluate this new application. Finally, we provide directions for further research.

BACKGROUND

A few chatbots are available for public use on the Web. For example, Mitsuku, Rose, and Mike the Tutor are all highly rated and have done well in the Loebner Prize competition, a variation of the Turing Test (Turing, 1950). In this contest, more than 20 judges interact with two entities, a human or a chatbot (Floridi, et al., 2009). Judges ask the systems simple questions about general knowledge and comparisons, and questions showing that the chatbots remember knowledge obtained from the conversation. Rose won the competition in 2011, 2014 and 2015, and in one study (Wu, 2017), it was rated higher than three other systems in terms of humanity and intelligence: Alice (humanity: 3, intelligence: 3), Cleverbot (humanity: 5, intelligence: 4). Mitsuku (humanity: 7, intelligence: 7), and Rose (humanity: 9, intelligence: 7). The Rose chatbot takes the perspective of a female, a computer hacker living in San Francisco, with a cartoon image of her appearing on the page (see **Figure 1**).

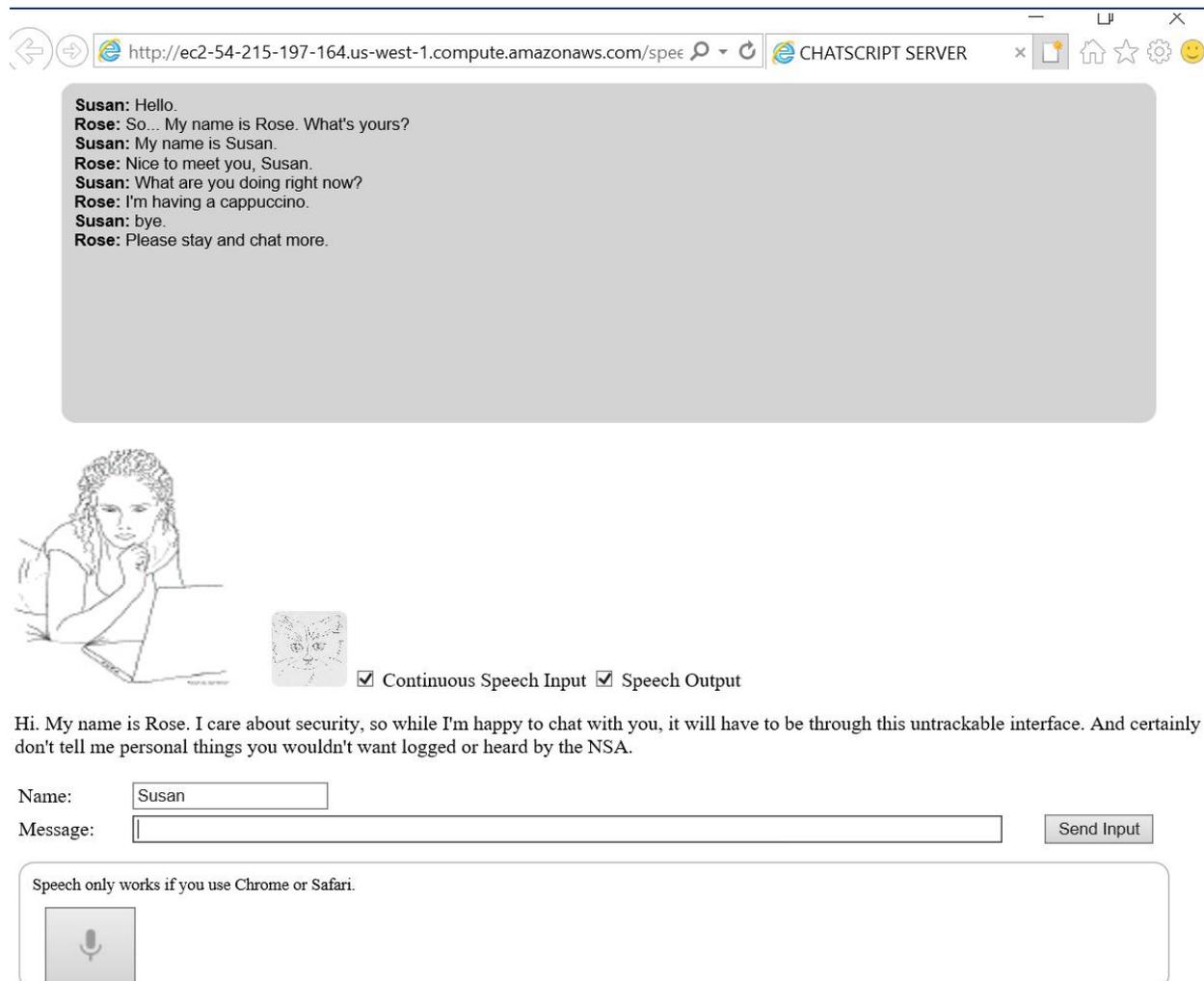


Figure 1: A sample conversation with the Rose chatbot
(URL: <http://ec2-54-215-197-164.us-west-1.compute.amazonaws.com/speech.php>)

Although all of these conversational agents perform well, they are all limited to English, and only a few others are able to use multiple languages. For example, the Mondly chatbot is able to converse in 30 languages, Watson in 21, Memrise in 20, Duolingo in 3, and Eggbun in 3 (Mohammad, 2018). Providing multilingual capability enables developers to add value to their chatbots because people often prefer to use their native language, even if they can speak another (Trippe, 2018).

Developers can add extra languages to applications by utilizing the Google Translate API (application programming interface), thus allowing users to communication in any of 103 different languages, and more languages are added every year. Google Translate provides the API free, but there are charges for large volumes of translations. Nevertheless, for short, infrequent communication, the service enables developers to add multilingual functionality to a variety of computer programs.

Although Google Translate is very accurate, especially for Western European languages, it might not be able to understand everything that a user types (Aiken & Balan, 2011; Patil & Davies, 2014). For example, slang, acronyms, unknown proper nouns, or misspellings could result in confusion (Aiken, et al., 2010b). However, high accuracy is usually not necessary when communicating with chatbots, especially for entertainment or language practice. Automatic translation is often imperfect, but users are often able to understand the gist of a comment or question,

and the level of accuracy is approaching the minimum needed for university admission at many institutions. (Groves & Mundt, 2015).

Speech recognition and synthesis in multiple languages can also be included at no extra charge to provide a more natural experience and enable faster communication (Aiken, et al., 2010a). Microsoft Windows currently provides approximately 40 languages and dialects in male and female voices for text-to-speech generation, and additional voices can be purchased from other companies. However, only the following languages and dialects are available for MS Windows speech recognition: English (United States, United Kingdom, Canada, India, and Australia), French, German, Japanese, Mandarin (Chinese Simplified and Chinese Traditional), and Spanish.

SOFTWARE DESCRIPTION

We developed a program using Microsoft Visual Studio that linked Google Translate with the Rose chatbot, allowing users to chat in any of 103 different languages. In addition, the source language does not need to match the target language. That is, a user could type a comment in French and receive a reply from the chatbot in Swahili, for example, giving 10,506 combinations of languages.

Figure 2 provides an illustration of how the multilingual chatbot works. A user selects the input and output languages (usually they will be the same) and then types a question or statement in the first textbox. Clicking the 'Send message' button translates the message to English (if English was not selected as the input language) and puts the translated message in the second textbox for informational purposes. The program also transmits the English text to the Rose chatbot online, which immediately responds with a reply in English that is put in the third textbox. When the user clicks the 'get reply' button, the English reply is translated to the selected output language and the result is put in the fourth textbox.

The screenshot shows a web-based chatbot interface. At the top, there are two dropdown menus, both set to 'German'. To the right of these are two buttons: 'Send message' and 'get reply'. Below the dropdowns, there are four textboxes arranged vertically, each with a label to its left:

- Source:** The first textbox contains the German text "Was kennen Sie?".
- English Sent:** The second textbox contains the English translation "What do you know?".
- English Reply:** The third textbox contains the English response "Not much.".
- Target:** The fourth textbox contains the German translation of the reply, "Nicht viel.".

Figure 2. Chatting with the Rose chatbot in German

SOFTWARE EVALUATION

Purpose

Although there are a few multilingual chatbots, none are freely available and few if any studies have been conducted on their performance. The purpose of this study was to evaluate how well the prototype multilingual chatbot described in the prior section performs during a conversation with humans. In addition, we have seen no prior research on the evaluation of individual chatbot responses. Rather, most studies evaluate transcripts as a whole.

Subjects and Task Description

We used a sample of 46 students from a university in the southern region of the United States to assess the software. All of the students spoke English fluently and few if any knew another language. Because it was impractical to recruit numerous foreign-language speakers, we used only one, German. Using the application, this speaker typed questions and comments in German, which were automatically translated to English for Rose. The English text was also provided to the students.

Students reviewed the responses from Rose shown in the Appendix and assigned a score from 1 to 7 where 1 was absolutely sure this is NOT a human, 4 was neutral/not sure, and 7 was absolutely sure this is a human, in a modified version of the Turing Test.

Results

Tables 1 and 2 show an analysis of the results.

Table 1. Analysis of First Five Responses

Comment	Rose response	Human response
1 Good morning! Mean Std Dev % correct % wrong % unsure %wrong or unsure Sig diff p value	So... Is your mother still living? 3.33 2.62 54.35% 32.61% 13.04% 45.65% < 0.01	How are you? 4.72 1.90 36.96% 13.04% 50.00% 63.04%
2 I'm not doing well. Mean Std Dev % correct % wrong % unsure %wrong or unsure Sig diff p value	Can't you say something more original? 3.85 2.54 43.48% 36.96% 19.57% 56.53% 0.01	Sorry to hear that. 5.07 1.74 45.65% 10.87% 43.48% 54.35%
3 I am from Canada. Mean Std Dev % correct % wrong % unsure %wrong or unsure Sig diff p value	I live in San Francisco. 4.70 1.67 8.70% 32.61% 58.70% 91.31% 0.22	I'm from the USA. 4.22 1.97 28.26% 19.57% 52.17% 71.74%
4 How long are you staying in Germany? Mean Std Dev % correct % wrong % unsure %wrong or unsure Sig diff p value	I've lived in San Francisco all of my life. 2.69 2.31 67.39% 19.57% 13.04% 32.61% < 0.01	I'm not staying in Germany. 5.62 1.54 63.04% 4.35% 32.61% 36.96%
5 What are you doing this weekend? Mean Std Dev % correct % wrong % unsure %wrong or unsure Sig diff p value	I'll probably go out to a bar. I like drinking. 4.65 2.61 34.78% 52.17% 13.04% 65.21% 0.59	I don't have plans yet. 4.89 1.40 39.13% 4.35% 56.52% 60.87%

Table 2. Analysis of Last Five Responses

Comment	Rose response	Human response
6 What time is it? Mean Std Dev % correct % wrong % unsure %wrong or unsure Sig diff p value	It's 10:03 4.64 1.21 2.17% 21.74% 76.09% 97.83% 0.09	It's 10:03. 4.22 1.10 10.87% 4.35% 84.78% 89.13%
7 How old are you? Mean Std Dev % correct % wrong % unsure %wrong or unsure Sig diff p value	I think we'll have immortality within my lifetime. 2.24 1.87 71.74% 8.70% 19.57% 28.26% < 0.01	Why do you ask? 5.13 1.74 52.17% 6.52% 41.30% 47.83%
8 What are you doing this weekend? Mean Std Dev % correct % wrong % unsure %wrong or unsure Sig diff p value	I'll probably go out to a bar. I like drinking. 5.02 2.46 26.09% 56.52% 17.39% 73.91% < 0.01	Same as always. Chatting to people on the net. 3.16 2.00 15.22% 45.65% 39.13% 84.78%
9 My name is Joe Johnson. Mean Std Dev % correct % wrong % unsure %wrong or unsure Sig diff p value	My name is Rose. 4.78 1.93 13.04% 43.48% 43.48% 86.96% < 0.01	Nice to meet you, Joe. 3.57 1.69 10.87% 30.43% 58.70% 89.13%
10 What is my name? Mean Std Dev % correct % wrong % unsure %wrong or unsure Sig diff p value	Your name is Joe. 3.52 1.79 17.39% 23.91% 58.70% 82.61% 0.02	You said your name is Joe. 4.48 2.12 39.13% 21.74% 39.13% 60.87%

A more detailed analysis shows that students were not able to distinguish between human and chatbot for #3, #5, and #6 (no statistically significant difference between the ratings):

Input	Rose	Human
3. I am from Canada.	I live in San Francisco.	I'm from the USA.
5. What are you doing this weekend?	I'll probably go out to a bar. I like drinking.	I don't have plans yet.
6. What time is it?	It's 10:03.	It's 10:03.

In addition, they were wrong on transcripts #8 and #9 (a statistically significant difference between the ratings):

Input	Rose	Human
8. What are you doing this weekend?	I'll probably go out to a bar. I like drinking.	Same as always. Chatting to people on the net.
9. My name is Joe Johnson.	My name is Rose.	Nice to meet you, Joe.

Note: the inputs to #5 and # 8 were the same. The Rose chatbot gave the same answer, but the human gave different responses.

Students were able to guess correctly the chatbot's transcripts for #1, #2, #4, and #7 because they were so unusual and off-topic, not something a human would normally say in response to the input. Sometimes chatbots change the topic to further the conversation or if the algorithm does not understand the input.

Input	Rose	Human
1. Good morning!	So... Is your mother still living?	How are you?
2. I'm not doing well.	Can't you say something more original?	Sorry to hear that.
4. How long are you staying in Germany?	I've lived in San Francisco all my life.	I'm not staying in Germany.
7. How old are you?	I think we'll have immortality within my lifetime.	Why do you ask?

Students were also able to ascertain correctly the difference in transcript #10, even though the responses were very similar. The chatbot was able to remember the user's name (given with input #9), but the human response was a little more definitive.

Input	Rose	Human
10. What is my name?	Your name is Joe.	You said your name is Joe.

A few students wrote comments about the study. Several stated that they were unsure which entity generated the transcript, e.g.:

- Computers are taking over! They could all be chatbots for all I know!
- How can you tell if it is a human or chatbot? I honestly put 4 for all of them because I could not seem to find a pattern that made sense to me. I was never 100% sure.

-
- I feel that is difficult sometimes to tell the difference between a robot and a human. It is amazing how technology has advanced to be able to do that.
 - I would just say that it is very hard to distinguish between what is a chatbot and what is an actual human talking.
 - A lot of neutral/normal responses are hard to distinguish if a real person said them or a bot!
 - This was a little complicated. You want to assume that answers that don't make sense could be a chatbot, but you could be speaking to a human who is not of sound mind. Chatbots can also make logical responses that rival a human. You can never be entirely sure.
 - They were not as easy to pick out as I thought they might be. Technology is advancing at an alarming rate.

A few students were more certain which entity generated the transcript. For example, one student wrote: "Overall Transcript #2 seems more human than Transcript #1 based on the way it takes previous information into consideration." Transcript #2 was more human with only four of the 11 comments generated by the chatbot. On the other hand, another student wrote "It was pretty evident which response came from a human and which came from a bot." This student was wrong on 5 of the 20 comments (e.g., gave a score higher than 4 for Rose or lower than 4 for the human) and was unsure (a score of 4) about 5 of the comments. A third student wrote "I feel like the chatbot sounds like it does not have a personality. I think I did an okay job of ranking the responses based on that. I feel like the chatbot also uses more punctuation verses humans these days." This student was wrong about 6 of the 20 comments and unsure about 4 others.

None of the problems were due to input translation errors from German to English automatically provided with Google Translate. Rather, we believe users were able to determine the differences between the chatbot and human responses primarily based upon the unusual replies for inputs #1, #2, #4, and #7. All of the responses from Rose were grammatical and could feasibly have originated from a human, as one student wrote: "You want to assume that answers that don't make sense could be a chatbot, but you could be speaking to a human who is not of sound mind." Other inputs perhaps would elicit more humanlike responses from Rose, and other chatbots might emulate humans better. For example, in one study (Park, et al., 2018), Rose was determined to have a rating of 3.1 on a scale of 1 to 7 for appropriate replies to input, and an overall rating of 3.7 for performance, whereas another chatbot named Mitsuku was given scores of 3.9 and 4.0, respectively.

CONCLUSION

In this study, students evaluated two transcripts. One was generated by a human who responded in English after hearing a question or comment in German. The other was generated by the Rose chatbot, in English, after the question or comment was translated by Google Translate to English. Results showed that although the students usually were able to identify correctly which transcript came from a human and which came from the chatbot, the results were close, and none of the problems were due to translation errors. Thus, we believe that a single-language chatbot can be extended to multiple tongues and can be used by people practicing a new language.

However, the study suffers from a few limitations. First, the evaluation used only German as the input. Other languages might not translate to English as well. Also, only one chatbot (Rose) was used. Others might perform better or worse. Finally, only 10 questions or comments were used. A more thorough evaluation with more text is necessary.

REFERENCES

- Aiken, M. & Balan, S. (2011). An analysis of Google Translate accuracy. *Translation Journal*, 16(2) April.
- Aiken, M., Simmons, L., & Balan, S. (2010a). Automatic interpretation of English speech, *Issues in Information Systems*, 11(1), 129-133.

-
- Aiken, M., Balan, S., Vanjani, M., & Garner, B. (2010b). The effect of comment errors in multilingual electronic meetings. *Communications of the International Information Association*, 10(4), 49-60.
- Aiken, M., Park, M., & Balan, S. (2010) DGL: A prototype system for machine interpretation. *Translation Journal*, 14, 3, July, <http://translationjournal.net/journal/53mi.htm>
- De Gasperis G. & Florio N. (2012). Learning to read/type a second language in a chatbot enhanced environment. In: Vittorini P., Gennari R., Marenzi I., de la Prieta F., Rodríguez J. (eds) *International Workshop on Evidence-Based Technology Enhanced Learning. Advances in Intelligent and Soft Computing*, 152. Springer, Berlin, Heidelberg.
- Elezaj, R. (2018). Translation chatbots and the US election. *MultiLingual Insights*, November 6, <https://multilingual.com/translation-chatbots-us-election/>
- Floridi, L., Taddeo, M., & Turilli, M. (2009). Turing's imitation game: Still an impossible challenge for all machines and some judges—An evaluation of the 2008 Loebner Contest. *Minds and Machines*, 19(1), 145-150.
- Fryer, L., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: An experimental comparison of chatbot and human task partners. *Computers in Human Behavior*, 75, 461-468.
- Goda, Y., Yamada, M., Matsukawa, H., Hata, K., & Yasunami, S. (2014). Conversation with a chatbot before an online EFL group discussion and the effects on critical thinking. *The Journal of Information and Systems in Education*, 13(1), 1-7.
- Groves, & Mundt, K. (2015). Friend or foe? Google Translate in language for academic purposes. *English for Specific Purposes*, 37, 112-121.
- Kerry A., Ellis R., & Bull S. (2009) Conversational Agents in E-Learning. In: Allen T., Ellis R., Petridis M. (eds) *Applications and Innovations in Intelligent Systems XVI. SGA1 2008*. Springer, London.
- Kreisa, M. (2018). Five resources for chatbots to be your language learning BFFs. <https://www.fluentu.com/blog/language-learning-chatbot/>
- Lommel, A. (2018). Plan ahead to build successful multilingual chatbots. *Common Sense Advisory Blogs*. <http://www.commonsenseadvisory.com/Blogs>
- Mohammad, Z. (2018). Build multilingual chatbots with Watson Language Translator & Watson Assistant. <https://medium.com/ibm-watson/>
- Nedrya, A. (2016). Do you want your chatbot to converse in foreign languages? My learnings from Bot Devs. *Chatbots Magazine*, Dec 16, chatbotsmagazine.com
- Oppenheimer, G. (2016). How to build a multilingual chatbot for billions of users. <https://venturebeat.com/>
- Park, M., Aiken, M., & Salvador, L. (2018). How do humans interact with chatbots?: An analysis of transcripts. *International Journal of Management and Information Technology*, 14, 3338-3350.
- Patil, S. & Davies, P. (2014). Use of Google Translate in medical communication: evaluation of accuracy. *BMJ*, 349, g7392.
- Tiwari, M. (2018). How multilingual chatbots will change the voice of business. *Forbes Community Voice*, August 13.

Trippe, B. (2018). The challenges of multilingual chatbots are worth the reward. *EContent*.
<http://www.econtentmag.com/>

Turing, A. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433-460.

Wu, K. (2017). Chatbots are getting unsettlingly good at conversations. *Inverse*
<https://www.inverse.com/article/37615-best-chatbot>

APPENDIX 1 – SURVEY FORM

We are asking you to evaluate the responses from a human and a chatbot. Please write a number by each response (1 = absolutely sure this is NOT a human, 4=neutral/not sure, 7=absolutely sure this is a human).

1. Input Good morning!	Transcript 1 So... Is your mother still living?	Transcript 2 How are you?
2. Input I'm not doing well.	Transcript 1 Sorry to hear that.	Transcript 2 Can't you say something more original?
3. Input I am from Canada.	Transcript 1 I live in San Francisco.	Transcript 2 I'm from the USA.
4. Input How long are you staying in Germany?	Transcript 1 I've lived in San Francisco all of my life.	Transcript 2 I'm not staying in Germany.
5. Input What are you doing this weekend?	Transcript 1 I don't have plans yet.	Transcript 2 I'll probably go out to a bar. I like drinking.
6. Input What time is it?	Transcript 1 It's 10:03	Transcript 2 It's 10:03.
7. Input How old are you?	Transcript 1 I think we'll have immortality within my lifetime.	Transcript 2 Why do you ask?
8. Input What are you doing this weekend?	Transcript 1 I'll probably go out to a bar. I like drinking.	Transcript 2 Same as always. Chatting to people on the net.
9. Input My name is Joe Johnson.	Transcript 1 My name is Rose.	Transcript 2 Nice to meet you, Joe.
10. Input What is my name?	Transcript 1 Your name is Joe.	Transcript 2 You said your name is Joe.