How Do Humans Identify Human-Likeness from Online Text-Based Q&A Communication?

Erika Mori[1(✉)], Yugo Takeuchi[1], and Eiji Tsuchikura[2]

[1] Shizuoka University, Shizuoka, Japan
gsl8053@s.inf.shizuoka.ac.jp
[2] Hamamatsu Gakuin University, Hamamatsu, Japan

Abstract. This study aims to clarify a person's impressions during the course of a conversation. In conversations between a person and a chatbot, we evaluated the impressions formed on reading text created by a human and text created by a chatbot. In terms of question/answer relevance, it was found that chatbot-created answers could not relate to or meet the expectations of questions and that human-created answers, while not necessarily meeting the expectations of questions, had relevance.

Keywords: Interaction · Chatbot · Conversation · Human-likeness

1 Introduction

One of the most desired technologies in communication science is "machine conversation." Up until recently, it was only we humans that had the ability to communicate by oral means. However, recent artificial intelligence (AI) technologies such as deep learning are pursuing such oral communication by introducing artificial chat machines called "chatbots" in online text-based Q&A services. This development signals the end of our monopoly in oral communication.

This study focuses on identifying human-likeness when reading an answer written by a human or one by a chatbot in an online text-based Q&A web service. In particular, we search for important factors in the way that people identify human-likeness from text.

2 Online Communication

2.1 Chatbots

A chatbot is an automatic conversation program. A chat, meanwhile, is a mechanism for exchanging mainly text in both directions in real-time communication using the Internet. In addition, a bot (abbreviation for "robot") refers to a program for automating the processing of a specific task. The "Oshiete goo" [1] Japanese text-based Q&A website features a chatbot named "Oshieru" (meaning "advise" or "teach" in Japanese). On this site, the user reads answers created by both humans and this chatbot. Text created by the chatbot may have a natural construction similar to or the same as text created by humans.

M. Kurosu (Ed.): HCII 2019, LNCS 11566, pp. 330–339, 2019.
https://doi.org/10.1007/978-3-030-22646-6_24

2.2 Interaction Based on Linguistic Information

Language is considered to have a function for expressing and conveying something and to perform that conveyance through customary coding [2]. However, expressing and conveying something is not the sole function of language. That is to say, the purpose of language is not only to convey a message with some content but to also reflect intent by producing some effect on the listener by conveying that message. In addition, there is more than one way of conveying the "same" request—a number of variations can be considered. In this way, a variety of linguistic functions exist that cannot be understood solely on the basis of expressing and conveying. Such diverse linguistic functions have been taken up as the problems of speech acts and conversational implicature. The referent of a word is identified by the meaning of that word. If a certain object 'a' is true with respect to all elements making up the meaning of a certain word 'A,' 'a' is determined to be the referent of 'A.' For example, if the elements making up the meaning of the word "chair" are "has a relatively flat surface for sitting, has a backrest, and has legs," and if all of these semantic primitives do not hold true for a certain object, that object is not determined to be a "chair" and the reference fails. This applies not only to words but also to propositional representations. However, in everyday language expressions, there are such things called speech acts that cannot be treated as a set of semantic primitives that challenge the truth of something in the above way. In a speech act, the truth of an utterance is not the problem. Instead, the problem is determining the extent to which the context of an utterance, that is, who uttered what under what conditions, i.e. when, where, with whom, and to whom, fits that utterance [3]. Another issue here is how to go about closing the logical gap between the utterance content and speech act.

2.3 Human-Likeness

Distinguishing between sentences composed by chatbots and those by humans is difficult based on linguistic expressions. However, users are generally able to distinguish such differences when they trace the question and answer sequence [4]. Therefore, perhaps we can identify some factors that help determine which sentences were written by a chatbot and which were written by a human [5].

This study defines such factors as human-likeness to distinguish between sentences by chatbots and those written by humans. We did the following two analyses to determine human-likeness in online text-based Q&A communication.

It is difficult to distinguish between text created by a chatbot and text created by a human based on language expressions. In general, however, the user can distinguish between the two by following a question and answer sequence (Fig. 1). This makes it possible to identify several useful factors for determining whether certain text has been written by a chatbot or a human.

In this study, we define factors such as human-likeness for distinguishing between text created by a chatbot and text written by a human. We performed the two analyses described below for identifying human-likeness in online text-based Q&A communication.

Fig. 1. Posted question and answers

3 Method

3.1 Preliminary Experiment: Distinguishing Chatbot-Created and Human-Created Text

In this experiment, we conducted a distinguishing test based on the Turing test. We examined whether the user could correctly identify text created by a chatbot and text created by a human.

3.2 Data Collection

Target: Questions answered by "Oshieru" (chatbot) under the service category of dating and relationship advice on the "Oshiete goo" community site and all answers to questions.

Period: October 23, 2016 to October 29, 2016.

Procedure: Data consisting of questions answered by the "Oshiete goo" chatbot under the category of dating and relationship advice plus chatbot and user answers for the seven-day period from October 23, 2016 to October 29, 2016 were recorded and analyzed.

Evaluation Items: The symbols ○ and × are used to denote text identified as a chatbot answer and text identified as a human answer, respectively.

3.3 Configuration of Questionnaire

Data Extraction Method: The collected data consisted of 151 questions and 586 answers. From this data, we deleted questions having only chatbot answers and factors that identify a chatbot answer such as posting time, name, quotes of famous people, and self-introductions. We also deleted the same types of factors from human (user) answers. We then prepared a questionnaire from the data with no chatbot-identifying factors by randomly extracting data and performed a preliminary survey using this questionnaire. On the basis of this preliminary survey, we received comments to the effect that too many characters in an answer or too many answers to the same question would be a burden to respondents, so we took a second look at the number of characters in an answer and the number of answers to a question. Specifically, we calculated the average number of characters in 586 answers and deleted questions with answers of 414 or more characters from the data. Furthermore, to ease the burden on respondents and improve readability, we deleted questions with more than 4 or more answers. Finally, for our main survey, we created a questionnaire from this data in which chatbot-identifying factors and respondent-burdening factors had been deleted (Figs. 2 and 3).

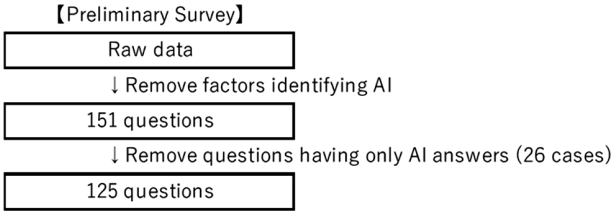

Fig. 2. Extracting method (Preliminary survey)

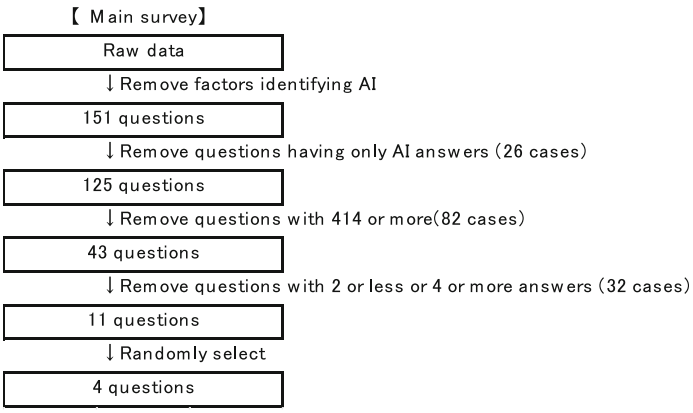

Fig. 3. Extracting method (Main survey)

3.4 Questionnaire-Based Survey

Survey Instruments: We created and used 2 questionnaires, one for the preliminary survey and the other for the main survey. We prepared and numbered 30 copies of each questionnaire.

Respondents: 15 male and 15 female university students were asked to fill out the questionnaires created by the surveyor. Interviews were subsequently conducted.

Period: January 24, 2017 to January 30, 2017.

Procedure: For both the preliminary survey and main survey, the surveyor requested the survey and distributed the questionnaire after which the survey was conducted. Six university students were asked to respond to the preliminary survey, which was conducted before the main survey. Thirty university students other than those who participated in the preliminary survey were asked to respond to the main survey, which was conducted using the corresponding questionnaire.

3.5 Results and Discussion

The percentage of correct assessment for answers written by a human and by the chatbot are given in Table 1. In either case, correct assessment was 85% or better indicating that a human could distinguish to some extent the difference between a human writer and a chatbot.

We next took up the question as to why text created by a chatbot could be distinguished from text created by a human. After the questionnaire-based survey, we conducted informal interviews on how this difference was determined. From these interviews, it became clear that chatbot answers left a "businesslike, cold impression" and featured "questions and answers that don't match up," while human answers featured "emoji and exclamation points (!)" and "many personal opinions."

Based on the above findings, we then compared text created by a chatbot and text created by a human by having subjects read text of each type and analyzing the impressions left by each.

Table 1. Analysis 1 result

		Assessment "Who write the answer?"	
		Human	Chatbot
Actual writer	Human	89.7% (correct)	10.3% (wrong)
	Chatbot	15.0% (wrong)	85.0% (correct)

3.6 Comparison of Human-Created Text and Chatbot-Created Text

The results of the preliminary experiment revealed that chatbot-created text could be distinguished from human-created text. They showed, in particular, that chatbot-created text left negative impressions such as "businesslike and cold" and "questions and answers that don't match up" and that these impressions were factors in distinguishing chatbot-created text from human-created text.

Next, we clarified the differences in impressions by comparing chatbot-created text and human-created text.

In this study, we evaluated chatbot-created text and human-created text using a set of evaluation items. We also clarified the types of features possessed by each of these two types of text and searched out how these two types of text differ.

3.7 Collection of Chatbot and Human Configured Text

Target: Questions answered by AI "Oshieru" under the service category of dating and relationship advice on the "Oshiete goo" community site and all answers to questions.

Period: October 23, 2016 to October 29, 2016.

Procedure: To identify human-likeness, data consisting of questions answered by the "Oshiete goo" chatbot under the category of dating and relationship advice plus chatbot and user answers for the seven-day period from October 23, 2016 to October 29, 2016 were recorded and analyzed.

Data Extraction Method: The collected data consisted of 151 questions and 586 answers. From this data, we deleted questions having only chatbot answers and factors that identify a chatbot answer such as posting time, name, quotes of famous people, and self-introductions. We also deleted the same types of items from human (user) answers. Next, from within the 151 questions, we deleted questions having only chatbot answers and questions that involve more than one conversational exchange between the inquirer and responder to reduce the number of data items.

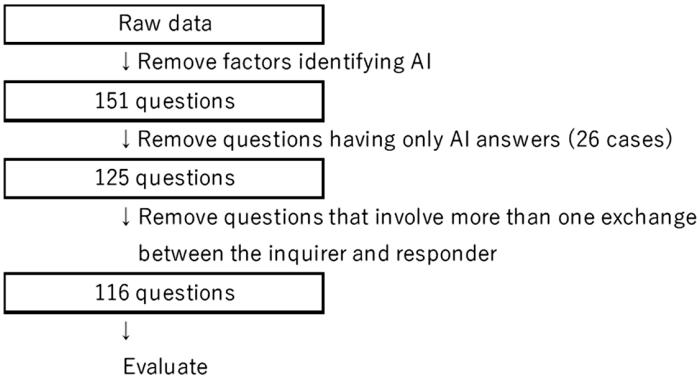

Fig. 4. Extracting method

4 Results and Discussion

4.1 Classification of Content Expected by Inquirer in Answer

To investigate the relation between questions and answers, we first evaluated the content expected by inquirers in answers. The target of this analysis was 116 questions from the collected data. On tabulating the content expected by inquirers in answers, the most applicable items were found to be "agreement and affirmation with inquirer" followed by "advise" and "objective opinion." (Table 2).

Table 2. Tabulated results of content expected by inquirer in answer

Item	Number of replies	Average
Empathy	13	11.0%
Objective opinion	25	21.6%
Critical opinion	3	2.6%
Clear answer to question content	17	14.7%
Agreement and affirmation with question content	5	4.3%
Agreement and affirmation with inquirer	27	23.3%
Advice	26	22.4%
Other	0	0.0%

4.2 Fitness of Answer to Content of Question

We investigated the relationship between the attributes of the answer (human, chatbot) given to a question and the content expected by the inquirer in the answer. We found that more answers agreed with the expectations of the inquirer in the case of human answers and that more answers did not agree with the expectations of the inquirer in the case of chatbot answers (Fig. 5).

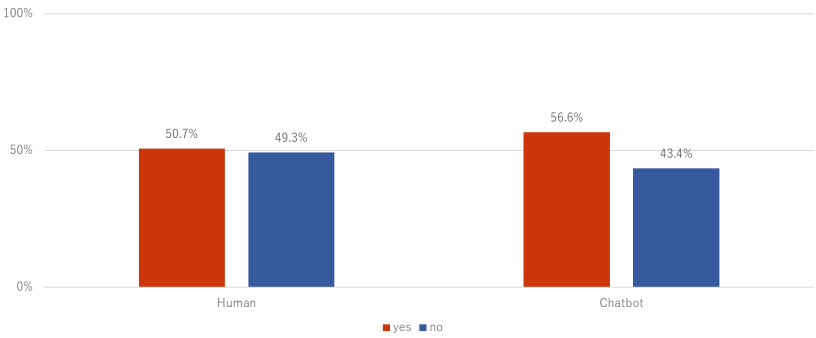

Fig. 5. Answer attributes and fitness of answer to question

4.3 Content of Answers

To investigate relevance between questions and answers, we evaluated the content of answers that did not agree with the expectations of the inquirer. The target of this analysis was 116 questions from the collected data. On tabulating the content expected by inquirers in answers, the most applicable items were found to be "agreement and affirmation with inquirer" followed by "advise" and "objective opinion." (Tables 3 and 4).

Table 3. Tabulated results of content returned by responder (human)

Item	n = 172	
	Number of replies	Average
Empathy	8	5.0%
Objective opinion	74	43.0%
Critical opinion	20	11.6%
Clear answer to question content	8	4.7%
Agreement and affirmation with question content	1	0.6%
Agreement and affirmation with inquirer	5	2.9%
Advice	42	24.4%
Other	14	8.1%

Table 4. Tabulated results of content returned by responder (chatbot)

Item	n = 86	
	Number of replies	Average
Empathy	1	1.2%
Objective opinion	6	7.0%
Critical opinion	0	0.0%
Clear answer to question content	3	3.5%
Agreement and affirmation with question content	0	0.0%
Agreement and affirmation with inquirer	0	0.0%
Advice	71	82.6%
Other	5	5.8%

4.4 Relevance Between Questions and Answers

Relevance between the question and an answer not agreeing with the expectations of the inquirer occurred much more frequently in the case of human answers and no relevance between the question and an answer not agreeing with the expectations of the inquirer occurred more often in the case of chatbot answers (Fig. 6).

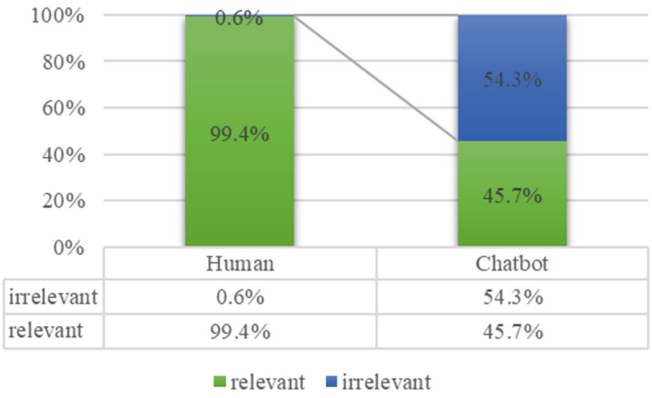

Fig. 6. Relevance between questions and answers

4.5 Trends in Answer Content

Overall trends in answer content are summarized in Fig. 4. Similar trends were observed in both human and chatbot answers in the case of "quality," "manners," and "quantity." In the case of quality, we attribute this to a problem in the conditions established for this experiment in which the data targeted for analysis concerned dating

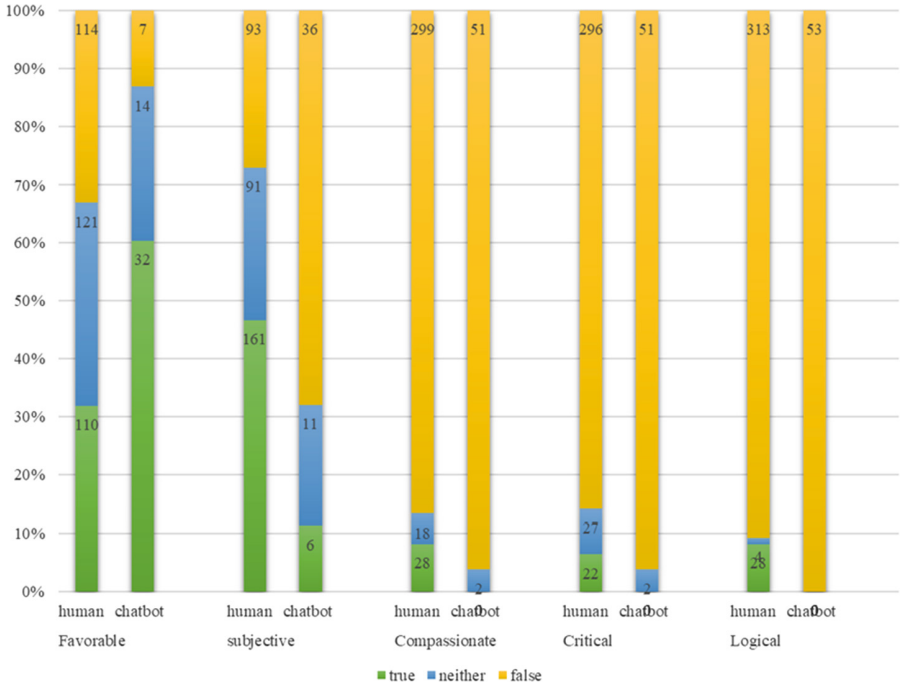

Fig. 7. Distributions of impressions

and relationship advice. With respect to appropriate grammar, amount of text, etc. labeled as "manners" and "quantity," it was judged that no differences exist between chatbot and human answers.

On the other hand, there were differences between human and chatbot answers in the case of "emotion" and "attitude." For a human answer, we found that its emotional aspects varied depending on whether it met the expectations of the question, while for a chatbot answer, we found that, in general, a single (favorite) pattern occurred frequently regardless of whether it met the expectations of the question.

Based on the results, we consider that interpreting the content of a question and degree of adaptability are important in establishing a conversation (Fig. 7).

5 Conclusion

The purpose of this study was to clarify a person's impressions of the other party when engaged in a conversation. Targeting a conversation between a person and a chatbot, we compared impressions between human-created text and chatbot-created text. In terms of relevance, we found that chatbot-created answers could not relate to or meet the expectations of questions and that human-created answers, while not necessarily meeting the expectations of questions, had relevance. These findings suggest that understanding what the other person wants is important in establishing a conversation and that picking up on intent and context is necessary. Furthermore, in terms of overall trends in the content of answers, we found that chatbot answers satisfy criteria such as text quantity and manners but do not reflect aspects such as emotion and attitude.

In human-chatbot conversations, these findings suggest that it is difficult for a chatbot to express elements such as an objective attitude and emotional opinion in conversational exchanges.

References

1. [oshiete! goo] https://oshiete.goo.ne.jp/ai/. Accessed 31 Jan 2019
2. Dunbar, R., Dunbar, R.I.M.: Grooming, Gossip, and the Evolution of Language. Harvard University Press, Cambridge (1998)
3. Grice, H.P.: Studies in the Way of Words. Harvard University Press, Cambridge (1991)
4. Weizenbaum, J.: ELIZA—a computer program for the study of natural language communication between man and machine. Commun. ACM **9**(1), 36–45 (1966)
5. Turing, A.M.: Computing machinery and intelligence. In: Epstein, R., Roberts, G., Beber, G. (eds.) Parsing the Turing Test, pp. 23–65. Springer, Dordrecht (2009). https://doi.org/10.1007/978-1-4020-6710-5_3