Answering What-type and Who-type Questions for Non-task-oriented Dialogue Agents

Makoto Koshinda, Michimasa Inaba and Kenichi Takahashi
Hiroshima City University, Hiroshima, Japan

Keywords: Non-task-oriented, Dialogue Agent, Question Answering.

Abstract: In this study, we propose a method for responding to what-type and who-type questions with no single fixed response handled by a non-taskoriented dialogue agent using Wikipedia as a language resource. The proposed method extracts nouns from a provided question text and then extracts an article from Wikipedia containing most of those nouns in its title. Next, words are extracted from the extracted article, and the degree of similarity between the extracted words and nouns extracted from question text is calculated. Words with a high degree of similarity are then acquired as response candidates. Next, the response candidates are ranked using the Wikipedia article structure and the text within the article, and the first-place response candidate is used for a response. According to the evaluation experiments, it was confirmed that the proposed method is capable of relatively natural responses in comparison to a baseline.

1 INTRODUCTION

In recent years, there has been vigorous research into dialogue agents, which often take the form of computers that converse with humans. Dialogue agents consist of two types: task-oriented dialogue agents that hold dialogue with people for the purpose of making ticket reservations or providing sightseeing information, for example, and non-taskoriented dialogue agents that do not perform a specific task and engage in daily conversation or chatting with users. Non-taskoriented dialogue agents are gaining attention not only for their use in entertainment, but they are expected to be used in various settings such as alleviating dementia and counseling (Sugiyama et al., 2013). Some known non-taskoriented dialogue agents include ELIZA (Weizenbaum, 1966), A.L.I.C.E (Wallace, 2008), and Microsofts Xiaoice [1]. Recently, there has been vigorous research into response generation by non-taskoriented dialogue agents using deep learning (Sordoni et al., 2015; Shang et al., 2015; Vinyals and Le, 2015). However, even with the use of deep learning, current non-taskoriented dialogue agents do not really have the same conversational capabilities as an actual person and thus require improvement in many areas.

One of these areas is responding to questions asked by users. The problem with past non-taskoriented dialogue agents is that they were often unable to properly respond to user questions and would fail in conversation. One possible method to solve this problem of failed conversation due to an inability to properly respond to the question is to set rules for questions and responses. However, when attempts are made to create rules for the most frequently occurring question text, very high costs are incurred Additionally, improved dialogue agent performance is reportedly limited even if rules are increased (Higashinaka et al., 2015). Therefore, in this study, we propose a method for properly responding to question texts with responses automatically acquired from Wikipedia. The method proposed in this study facilitates proper responses by using the similarity of response candidates extracted from Wikipedia articles to question text, and the structure and text of articles for which titles match response candidates.

2 RELATED WORK

Several types of questions exist: what-type questions about objects like “What game do you like?” who-type questions asking people “Who do you respect?” why-type questions asking for reasons like “Why do you like tomatoes?” and how-type questions asking about means or methods like “How do you take good

[1] http://www.msxiaoice.com/

Koshinda, M., Inaba, M. and Takahashi, K.
Answering What-type and Who-type Questions for Non-task-oriented Dialogue Agents.
DOI: 10.5220/0006517102870294
In *Proceedings of the 10th International Conference on Agents and Artificial Intelligence (ICAART 2018) - Volume 2*, pages 287-294
ISBN: 978-989-758-275-2

pictures?" Studies have been conducted on how to properly respond to these questions.

For example, a study by Sasaki et al. (Sasaki and Fujii, 2014) researched responses to how-type questions. This study proposed a method for solving problems with existing methods, such as selecting responses that commonly appear for all questions, or selecting "Ask someone else" as a response. Specifically, it proposed a method in which a document search is conducted of question text, and the inverse document frequency of the documents appearing as search results is calculated. This results in a reduced search score for response candidates such as "Ask someone else," which are likely to appear as the response to many questions, or as a lack of response to a specific question. Based on the results of comparing existing methods and their performance, the highest percentage of correct responses appears for the top 100 response candidates. Additionally, they report that based on the results of verifying the effectiveness of multiple proposed methods for solving the problems in existing methods, combining all of the proposed methods produces better results than other methods.

In a study of why-type questions, Fukumoto et al. proposed a method using manually created expression patterns of "because of" and "by reason of" (Fukumoto, 2007). However, the problem here is that when an attempt is made to extract response candidates using such manually created patterns, it is not possible to accommodate keywords other than these patterns (ex: "for the sake of . . ." etc.), and when an attempt is made to accommodate many key phrases manually, the cost of creating patterns increases. Therefore, Higashinaka et al. proposes a corpus-based method for reducing dependency upon manually created key phrases (Higashinaka and Isozaki, 2008). In this method, sections tagged with a "cause" within a corpus are extracted, and then cosine similarity is used to calculate the similarity of question text and response candidates. Then, the response candidates are ranked according to prepared characteristics, and then a response is given. According to the results of comparing multiple methods, including this one, compared to other methods, the results show a high percentage of questions that may be responded to within the top nth positions. In addition, of the several prepared characteristics, the characteristics prepared using the "cause" tag and the similarity characteristics are reportedly bound to the functions of the proposed methods in particular. Verberne et al. proposes a system for responding to why-type questions using Wikipedia as a response resource (Verberne et al., 2010). In this system, performance is improved in comparison to existing question and response systems as a result of considering key words such as "because" and words overlapping with the Wikipedia articles serving as a response resource. Additionally, Verberne et al. report improved ranking performance as a result of considering the structure of articles, such as the titles of Wikipedia articles, which suggests the effectiveness of using Wikipedia as a response resource.

Other than this study, there are other studies (Clarke et al., 2010) of what-type and who-type questions by Clarke et al. They use pre-provided semantic information in question texts such as "What is the largest state that borders Texas?" and data from a set of responses to question text for learning purposes. Then, they predict semantic information in question texts inputted from results and determine the responses based on the semantic information. Furthermore, a method using a recurrent neural network (RNN) (Iyyer et al., 2014) has recently been proposed. However, these studies are of a single response to a question. In chatting, there are many questions with different responses depending upon the person, such as "What foods do you like?" or "What sports do you like?" Therefore, in the present study, we propose a method for responding to "What" and "Who" questions in which a single response is not decided for a question.

3 APPROACHES

The proposed method provides proper word output as a response to arbitrary what-type or who-type questions. For example, if the question "What sports do you like?" is asked, the response may be "Tennis" or "Soccer."

The following is a rough procedure of the proposed method:

1. Extract words from question text and acquire article from Wikipedia
2. Extract words with hyperlinks to articles different from acquired article
3. Calculate degree of similarity between words extracted from words in question text and article
4. Based on the results in 3), use 10 words close in meaning to words in question text as response candidates and extract an article from Wikipedia in which the article title consists of response candidates.
5. Search for text containing many question text words in articles with titles consisting of response

candidates and then weigh the response candidates according to where in the article the text is located.

6. Calculate the degree of similarity between text in article used for weighing in 5) and question text
7. Score the product of the weight calculated in 5) and the degree of similarity in 6), rank the response candidates according to this score, and then respond with the first-place response candidate

The ranking of words extracted from Wikipedia through such a procedure makes it possible to acquire words thought to be close in meaning to question text by filtering them by the degree of similarity between words in question text and extracted words. Next, as a result of filtering, words close in meaning to question text are used as response candidates, and then an article in which the title consists of response candidates is extracted from Wikipedia. If text in an extracted article contains many of the words in question text, then response candidates containing this text within the article are relevant to question text; furthermore, it is felt that the relevancy is even greater if the text resembles question text.

3.1 Extraction of Response Candidates using Word2Vec

In the proposed method, response candidates are extracted according to the degree of similarity between words in a Wikipedia article and the nouns in question text. Thus, it is possible to acquire from Wikipedia words that are close in meaning to words in question textin other words, words that may serve as responses. For example, when the question text is “What sports do you like?” words are acquired from an extracted article in response to the noun “sports” in the question text. For example, “tennis,” “soccer,” “Italy” and “Internet” are words that may be called up for their relevance to “sports,” and then, in this case, “tennis” and “soccer” would be acquired as response candidates. To calculate the degree of similarity between words, Word2Vec (Mikolov et al., 2013b; Mikolov et al., 2013a) proposed by Mikolov et al. is used in the present study.

First, all nouns are extracted from initially provided question text. Next, using the words inside of this extracted question text, the article containing most of these words in its title is obtained from Wikipedia. Then, all words with hyperlinks to other articles in Wikipedia are extracted from the article. Next, all word vectors inside question text vectorized with Word2Vec are summed up and the respective cosine degrees of similarity of these vectors and each word vector within Wikipedia are calculated. Lastly, *n*th words within Wikipedia in order of highest cosine degree of similarity, that is, the ones considered the most similar to words within question text, are taken to be response candidates. In evaluation experiments, *n* is set to be 10.

3.2 Weighing of Response Candidates using Wikipedia Article Structure

In this section, we will explain a method in which a Wikipedia article is obtained with a title that consists of response candidates acquired in the preceding section, and then the article structure is used to weigh the response candidates.

A Wikipedia article first has a lexical definition that is the subject of the article, followed by a summary, and then a detailed explanation of the origin or history of the subject. Hence, if there is a text containing many words in question text in the definition and summary, then the relevance between the words and question text is considered high. Therefore, in this study, we always use text in articles close to the top that also contain the most words within question text and then find the weight W of each response candidate with Formula (3.2) below.

$$W = (a(1-\frac{pos_{sec}}{N_{pos_{sec}}}))(b(1-\frac{pos_{sent}}{N_{pos_{sent}}}))(c(1-\frac{pos_{all}}{N_{pos_{all}}})) \quad (1)$$

The symbols pos_{sec}, pos_{sent} and pos_{all} represent at what number a section containing text contains the most words from question text appearing within an article, at what line it appears within a section, and at what line it appears in the article overall. In addition, $N_{pos_{sec}}$, $N_{pos_{sent}} and N_{pos_{all}}$ respectively represent the total number of sections within an article, the total number of lines within a section, and the total number of lines within an article. That is, W is large in value when a section containing a text containing the most words within question text is at the top of the article, and furthermore, the text is at the start of a sentence within the section and in the article overall. Further, a, b, and c are respective parameters for setting importance, and in the experiment described below, we used $a = 1.5$, $b = 1.2$ and $c = 0.8$. These values were empirically determined.

3.3 Similarity between Sentences in Wikipedia Article and Question Text

For response candidates weighed in section 3.2, we calculate the degree of similarity of the text that is the basis for weighing in Formula (3.2), that is, the text containing the most words within question text and provided question text. If the degree of similarity is high, it is because the text is close in meaning, and the relevance of the response candidates to question text is considered higher. If a list of the nouns respectively appearing in each of the texts is created for this calculation, then text vectors are created: "1" if the words in the list are used in the text and "0" if the words in the list are not used in text. Then, using these vectors, the cosine similarity is calculated for the degree of similarity between the baseline text and question text. For example, nouns "A" and "B" are contained in question text, and nouns "A," "C," and "D" are contained in the text in Wikipedia containing the most words in the question text. When this happens, the question text vector is $v_q = (1,1,0,0)$, the Wikipedia text vector is $v_w = (1,0,1,1)$, and the cosine similarity of the two vectors is the degree of similarity of the text.

The degree of similarity s_{v_q,v_w} of the text thus derived is applied to the weight W found using Formula (3.2), and a score is calculated using Formula (2).

$$score = W \times s_{v_q,v_w} \tag{2}$$

4 EXPERIMENT

4.1 Settings

In this experiment, the proposed method was manually evaluated. A total of 10 evaluators each evaluated the suitability of response candidates for 100 lines of question text. The evaluators were recruited on crowd-sourcing site CrowdWorks[2].

In addition, for the 100 lines of question text used in this experiment, a request for question text to be prepared was made with CrowdWorks separately from the experiment, which was then collected. Table 1 shows an example of the questions used for evaluation.

For subjects, multiple response candidates were provided for one line of question text, and then an evaluation was provided on the following three-point scale:

G (Good). Response candidates are suitable question text responses

A (Average). Response candidates are not necessarily unsuitable as question text responses, but they feel awkward

P (Poor). Unsuitable as question text responses

Table 1: Example of question text evaluated in an experiment.

What movie do you like?
What is your favorite animal?
What is your most recent impulse purchase?
What games do you play?
Which politicians do you respect?
What TV programs do you frequently watch?
What snacks do you like with your drinks?
What sports do you do?
What job do you do?
Any modern artists you like?

4.2 Comparing Methods

In this experiment, we used three methods: word degree of similarity, article structure, and article structure + degree of similarity for comparison.

4.2.1 Word Degree of Similarity

The word vectors described in section 3.1 were used for word degree of similarity, calculating the degree of similarity between words extracted from a Wikipedia article and words extracted from question text with the cosine similarity. With this method, response candidates were ranked from largest to smallest degree of similarity.

4.2.2 Article Structure

The degree of similarity was calculated between words within question text described in Section 3.1 and words extracted from a Wikipedia article, and then 10 words extracted from the Wikipedia article with a high degree of similarity to words in question text were used as response candidates. With this method, an article was acquired with a title consisting of response candidates described in Section 3.2, then the article structure was used to weigh the response candidates according to where they were located in the article text containing many words from question text, and then they were ranked from the largest to smallest value of weight W.

4.2.3 Article Structure + Degree of Similarity

Article structure + degree of similarity is a method combining all of the methods proposed in Chapter 3. Specifically, for response candidates weighed using the article structure described in Section 3, the degree of similarity was calculated between texts containing many words in question text used for weighing and question text. The weighed value was applied to the

[2] https://crowdworks.jp/

Table 2: Accuracy rate of each method.

Method	G	G + A
Word degree of similarity	0.41	0.63
Article structure	0.43	0.67
Article structure + degree of similarity	**0.47**	**0.68**

degree of similarity, and then they were weighed from largest to smallest score.

4.3 Evaluation Method

We evaluated each method by the percentage of proper responses in first place (accuracy rate) and mean average precision (MAP).

First, with the percentage of proper responses in first place (accuracy rate), we assume it was possible to properly respond to questions if response candidates that had been assessed as accurate by the majority of subjects were ranked in first place.

MAP is an index indicating how much accurate response candidates are appearing in the top *n*th of ranked tasks. In this experiment, similar to accuracy rate, if response candidates were assessed as accurate by a majority of subjects, then the response candidates were considered accurate, and we calculated MAP.

5 RESULTS

Table 2 shows the results of calculating the accuracy rate of each method. According to the results in Table 2, the accuracy rates when only G was considered accurate, and when both G and A were considered accurate, confirmed that a combination of all the methods proposed in Chapter 3 had the highest accuracy rate.

Next, Fig. 1 and 2 show the respective results of each method when rank *n* during calculation of MAP was changed to first to tenth place. Furthermore, Fig. 1 shows the results when only G was considered accurate, and Fig. 2 shows the results when both G and A were considered accurate.

Based on the above, when only G was considered accurate, and when both G and A were considered accurate, the best results other than first place were when using article structure. First, from the results in Fig. 1 and 2, article structure was often used when accurate response candidates were ranked at the top from second place on down.

Based on this, we found that using a Wikipedia article structure to weigh where texts containing many words in question text are appearing within an article with a title consisting of response candidates is effective in raising the rank of appropriate response candidates to question text. Additionally, we found that the degree of similarity between question texts and texts containing many words in question texts serving as the basis for weighing should be considered in selecting the most appropriate response to question text from among the top response candidates listed by using Wikipedia article structure. However, the results in Fig.1 and 2 confirm that the MAP is larger for article structure than article structure + degree of similarity in nonfirst place cases. These results are discussed in the following section.

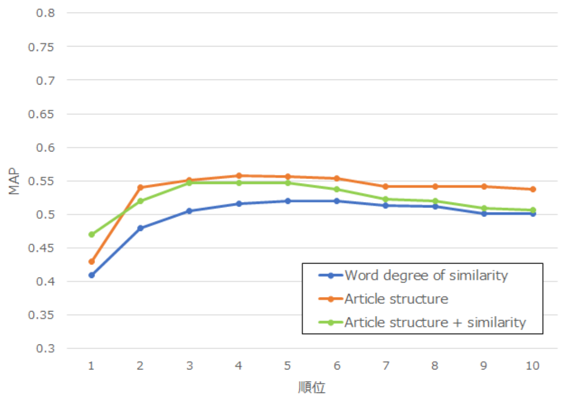

Figure 1: MAP when G is considered accurate.

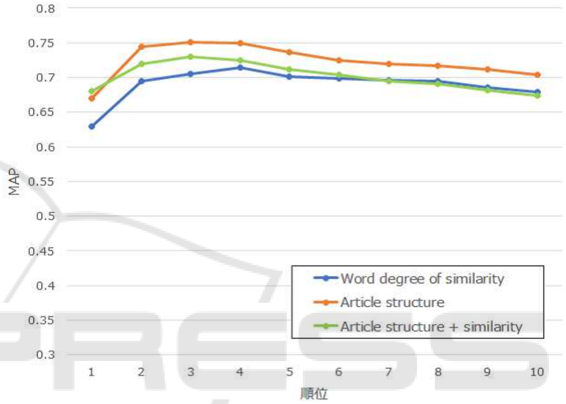

Figure 2: MAP when G and A are considered accurate.

5.1 Discussion

The results in Table 2 confirm that article structure + degree of similarity is better than using the ranking of words for which only the degree of similarity between words extracted from an article and words within question text have been extracted. We shall compare and discuss the results obtained as to why it was possible to raise the accuracy rate in this manner. Table 3 shows one set of the results obtained.

Table 3: Results of ranking the question, “What was the first toy you ever bought?”.

Rank	Word degree of similarity	Article structure + degree of similarity
1	Toy	Educational toy
2	Educational toy	Plush doll
3	Plush doll	Model train
4	Lego	Plarail
5	Tomica	Building blocks
6	Model train	Transformation belt
7	Plarail	Furniture
8	Building blocks	Dress-up doll
9	Transformation belt	Stationary
10	Furniture	Nanoblock

Table 4: Results of ranking the question, “You want to change to a single lens reflex camera, but which manufacturers are recommended?”.

Rank	Word degree of similarity	Article structure + degree of similarity
1	Nikon	Video camera
2	Single lens reflex camera	Rodenstock
3	Camera	Camera
4	Digital camera	Wide angle lens
5	Pentaflex	Pentflex
6	Epson	Single lens reflex camera
7	Rodenstock	Nikon
8	Speedlite	Digital camera
9	Wide angle lens	Mirror single lens reflex camera
10	Video camera	Epiphone

From the results in Table 3, under word degree of similarity, “Toy” is in first place, while under article structure + degree of similarity, “Educational toy” is in first place. Such results are also seen with questions such as “Previously, what sort of diets have you tried?” In this example, under word degree of similarity, “Reducing weight” was in first place, while under article structure + degree of similarity, it was “Low sugar diet.” Hence, under word degree of similarity, a word indicating the same meaning as a word contained in question text is in first place, while under article structure + degree of similarity, a response candidate related to words in question text is in first place. From this, it is felt that the relevancy of response candidates and question text can be considered by using Wikipedia article structure, then a search can be made for text containing a lot of words in question text from an article with a title consisting of response candidates, and then the degree of similarity of this text and question text can be calculated.

Additionally, from the results in Fig. 1 and 2, methods with the largest MAP other than first place were ones that carried out weighing using the article structure in Section 3.2, instead of article structure + degree of similarity. Therefore, we surveyed response candidate articles and the results of ranking question text, which tended to be higher with methods that carried out weighing in which article structure was used for the MAP of obtained questions.

The results of ranking obtained with the question, “You want to change to a single lens reflex camera, but which manufacturers are recommended?” are shown in Table 4. Among the response candidates, those in first place were from methods using article structure, while we found articles with the response candidate “Nikon” in sixth place with methods using article structure + degree of similarity. As a result, we found that in an article on “Nikon,” texts containing the most nouns in question text had a long text length and contained many nouns. That is, when a vector was made to calculate the degree of similarity of text between the article text and question text, the Wikipedia text vector was large in dimensionality, and most of its elements were 1. The question text vector becomes a vector in which most of its elements are 0. Consequently, we found the degree of similarity between these two vectors to be small. Otherwise, the same trends were seen in the results for question text “What to you is a scary disaster?” and the results for other question text. Namely, if the number of nouns contained in Wikipedia text used to calculate the degree of similarity between texts is large, then the degree of similarity is inversely proportional and tends to be small.

However, from the results in Table 2, high accuracy rate results were obtained when the cosine degree of similarity was calculated between texts containing many question text words from an article with a title consisting of response candidates and question text. Accordingly, it is felt that calculating degree of similarity between text containing many words in question text from a Wikipedia article and question text is effective in responding with an appropriate response. However, since there are cases in which this has a negative impact, as in this example, calculation of degree of similarity requires study of methods for compensating for the impact of the number of words in future.

6 CONCLUSION

In this study, we proposed a method for responding to what-type questions like “What sports do you like?” and who-type questions like “Who do you respect?” using Wikipedia as a language resource. The proposed method first extracted words from question text and then extracted an article containing most of these words in its title from Wikipedia. Next, words were extracted from the extracted article, the degree of similarity between the extracted words and words extracted from question text was calculated, and then words extracted from the article thought closest in meaning to the words in question text were acquired as response candidates. Furthermore, an article with a title consisting of response candidate words was acquired from Wikipedia and, using the structure of the obtained article, the response candidates were weighed. Then, the degree of similarity between texts containing many words in question text in the article with a title consisting of response candidates and question text was calculated, values applying weights calculated using article structure to the degree of similarity were used as scores for response candidates, and a response was made with the first-place response candidates. We found that a method combining all of the proposed methods, compared to other compared methods, can respond to questions with a high accuracy rate. Additionally, the method is effective for considering the use of Wikipedia article structure and the degree of similarity between article text and question text.

As part of future work, we would like to review methods of calculating the degree of similarity between article text and question text and evaluate the proposed method by applying it to a non-taskoriented dialogue agent and conducting actual dialogues with a person.

ACKNOWLEDGEMENTS

This study received a grant of JSPS Grants-in-aid for Scientific Research 16H05880.

REFERENCES

Clarke, J., Goldwasser, D., Chang, M.-W., and Roth, D. (2010). Driving semantic parsing from the world’s response. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 18–27. Association for Computational Linguistics.

Fukumoto, J.-i. (2007). Question answering system for non-factoid type questions and automatic evaluation based on be method. In *NTCIR*, pages 441–447.

Higashinaka, R. and Isozaki, H. (2008). Corpus-based question answering for why-questions. In *IJCNLP*, pages 418–425.

Higashinaka, R., Meguro, T., Sugiyama, H., Makino, T., and Matsuo, Y. (2015). On the difficulty of improving hand-crafted rules in chat-oriented dialogue systems. In *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pages 1014–1018.

Iyyer, M., Boyd-Graber, J. L., Claudino, L. M. B., Socher, R., and Daumé III, H. (2014). A neural network for factoid question answering over paragraphs. In *EMNLP*, pages 633–644.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Sasaki, S. and Fujii, A. (2014). Enhancing how-type question answering based on predicate-argument relations. *IPSJ Journal*, 55(4):1438–1451.

Shang, L., Lu, Z., and Li, H. (2015). Neural Responding Machine for Short Text Conversation. *Proceedings of the 53th Annual Meeting of Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1577–1586.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., and Dolan, B. (2015). A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the NAACL-HLT 2015*, pages 196–205.

Sugiyama, H., Meguro, T., Higashinaka, R., and Minami, Y. (2013). Open-domain utterance generation for conversational dialogue systems using web-scale dependency structures. In *Proc. SIGDIAL*, pages 334–338.

Verberne, S., Boves, L., Oostdijk, N., and Coppen, P.-A. (2010). What is not in the bag of words for why-qa? *Computational Linguistics*, 36(2):229–245.

Vinyals, O. and Le, Q. (2015). A neural conversational model. In *Proceedings of the ICML Deep Learning Workshop*, pages 1–7.

Wallace, R. (2008). The anatomy of A.L.I.C.E. *Parsing the Turing Test*, pages 181–210.

Weizenbaum, J. (1966). ELIZA-a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.