An Automatic Question Answering Method for Small-Scale Corpus

Yikang Shao¹, Guangde Xu¹,², Mingxue Xu¹ and Lanfang Dong¹

¹School of Computer Science and Technology, University of Science and Technology of China, Hefei, China
²Guochuang Software Co.Ltd., Hefei, China
Email: gdxu@mail.ustc.edu.cn

Abstract. It is a growing trend for automatic question answering system to be prominent in the development process of society. There are many methods trying to address this problem, but with deficiencies—relatively developed methods based on template matching need a lot of manual work writing templates, and machine learning based methods need plenty of work collecting a large number of corpus, bring huge burden on small-scale scenery. Facing these problems, we propose an automatic question answering methods meeting the needs of small-scale corpus. This method consists of combining an improved text similarity calculation algorithm and an intention recognition method based on slot filling. We conduct experiments on the problem sets of related fields, and it shows a good performance of the proposed automatic question answering method. Our methods make small scene applications for dialog systems more practical.

1. Introduction

With the rapid development of artificial intelligence, automatic question answering systems have been widely used in some online customer service scenarios. However, the development of this type of automatic question answering system often requires a large-scale corpus as a support. It is difficult to quickly implement a targeted dialogue system when facing a small-scale corpus scene. For some traditional question-and-answer fields that have not yet been informatized, it is difficult to obtain large-scale corpus data in the corresponding dialog system research. Therefore, how to develop an automatic question answering system for a small-scale corpus is an important issue today.

Although the traditional automatic question answering method based on template matching does not require large-scale corpus as a support, it needs to manually write a large number of templates for each type of problem, which is time-consuming and labor-intensive, and has low efficiency. Even if the corpus is small, it is difficult to complete the rapid development and deployment of the dialogue system.

To solve the above problems, we propose an automatic question answering method based on the combination of text similarity and intention recognition for small-scale corpus. In this method, the single-round dialogue strategy based on text similarity uses an improved text similarity algorithm, while the multi-round dialogue strategy based on intent recognition draws on the idea of template matching. Aiming at the target question raised by the user, the results of the two dialogue strategies are weighed and compared, and corresponding answers are given.

The remainder of the paper is organized as follows. Section 2 reviews the research progress of the automatic question answering method; Section 3 details the automatic question answering method for
small-scale corpora; Section 4 gives the experimental evaluation of the automatic question answering method; Section 5 summarizes this article and gives the direction of the next step.

2. Related Work
The research on automatic question answering systems has a history of decades. Commonly used automatic question answering methods can be divided into methods based on template matching and methods based on machine learning.

2.1. Automatic Question Answering Method Based on Template Matching
In the 1950s, Turing, the founder of computer science, first proposed the concepts of “machine intelligence” and “Turing test” [1]. Since “Turing test” adopts the form of human-machine dialogue, it is also widely regarded as the originator of the automatic question answering system. In 1966, Joseph of the Massachusetts Institute of Technology developed the earliest practical chatting system Eliza [2]. Eliza was originally developed primarily for the medical field, but it is actually an extensible scripting engine. However, due to the limitations of its script design, Eliza did not perform well in practical applications.

In 1995, Dr. Richard S. Wallace proposed the artificial intelligence markup language AIML, and on this basis developed the artificial intelligence chat robot A.L.I.C.E. [3]. AIML is an XML grammar that can be used to write rules for automatic question answering. A.L.I.C.E. has a simple design but is very effective. It has won three Lebner Prizes and the 2004 Chatterbox Challenge competition [4]. A.L.I.C.E. is also an automatic question answering system platform that can be used in a chat system in multiple languages and multiple fields [5].

In 1998, Wendell developed Talk-Bot by JavaScript and PERL language, which is an online chat system [6]. Later, Bruce conducted research on conversation management with context in mind, and developed the automatic question answering system engine ChatScript. The robot won the Lebner Prize in 2010, 2011 and 2014 [7].

The automatic question answering method based on template matching is getting better and better, but they are based on a lot of human work to write the corresponding template, which takes time and effort, and cannot meet the needs of rapid development.

2.2. Automatic Question Answering Method Based on Machine Learning
With the development of machine learning technology in recent years, automatic question answering methods based on machine learning are also on the rise. The automatic question answering method based on machine learning can be regarded as a text classification method in essence, and it can be roughly divided into a generative automatic question answering method and a retrieval automatic question answering method.

The generative automatic question answering method draws on the idea of machine translation, mainly by building a neural network and training it with a large amount of corpus. The trained model can automatically generate corresponding answers to user questions. The idea of Seq2seq (sequence-to-sequence), a commonly used generative automatic question answering framework, was first proposed by Cho in 2014 [8]. This framework was first used for machine translation, and it is a kind of Encoder and Decoder. Variant of the recurrent neural network (RNN). Although it does not require excessive manual intervention, the effect of generating answers is often not very satisfactory [9].

The common idea of the searchable automatic question answering method is to quantify the target question, and select the appropriate answer according to the vector matching between the target question and the candidate question [10]. In 2015, Feng et al. conducted an experimental test on the TF-IDF algorithm and neural networks with various structures using an insurance industry corpus in IBM Watson’s research, and found that the CNN-based vector construction method works well [11]. In the research field of Chinese automatic question answering system, Li et al. and Zhou et al. have proposed an improved algorithm for the calculation of Chinese text similarity [12, 13].
In summary, in the development of automatic question answering methods, the method based on template matching has tended to mature, but often only for open question answering, or requires a lot of manual writing. The template is inconvenient for the development of automatic question answering systems for small-scale corpora; machine-based methods rely on statistical rules, so they have high requirements for the size of the corpus, and they often do not get good results when developing for small-scale corpora.

3. Automatic Question Answering Method Based on Combination of Text Similarity Calculation and Intention Recognition

The two automatic question answering methods based on template matching and machine learning are introduced above, and their respective problems are analyzed separately. This article is mainly for small-scale corpus dialogue systems with a total number of corpora of less than 1,000. When the corpus size is small, the template-based method will be difficult to include all kinds of variants of the problem in this field, based on machine learning. The method cannot learn complete and accurate language laws. In this context, this section will introduce a new automatic question answering method, which is a combination of a single-round dialogue strategy based on an improved text similarity algorithm and a multi-round dialogue strategy based on intent recognition. This section introduces the framework and specific algorithm of the automatic question answering method, and gives the corresponding tests and experiments in the next section to illustrate the effectiveness of the method.

3.1. Overall Framework of Automatic Question Answering Method

The overall automatic question framework of answering method for small-scale corpus is shown in figure 1.

![Figure 1. Framework of question answering method.](image)

It can be trimmed from figure 1 that when a target problem is input into the system, the system first transforms the text segmentation, replacement word removal and other preprocessing tasks, and then calculates the text similarity between the target problem and the corpus reduction problem one by one. When the text similarity of a specific question reaches a certain threshold, the system processes it according to a single round of dialogue strategy and directly outputs the answer corresponding to the indicated question. Then the target question is identified according to Threshold A. If it is correctly identified, it will be processed according to multiple rounds of dialogue strategy, find the corresponding alternative question and output its corresponding answer. The mean value of the text similarity of the alias problem is less than the threshold value A, and it must be recognized that the recognition is unsuccessful, then continue to determine and request the user to re-enter.
3.2. Improved Text Similarity Algorithm Based on Synonymous Word Forest

The commonly used text similarity algorithm is a method based on TF-IDF, this method first composes the target problem and all candidate problems into a problem set, and builds a bag-of-words model that contains all the words in the problem set, and then calculates the TF-IDF value of the words one by one, and then calculates each The TF-IDF weighted one-hot vector of the problem. Finally, the cosine similarity between the vector corresponding to the target problem and the candidate problem is calculated to describe the text similarity between them.

In the traditional method, each different word corresponds to a single dimension in the TF-IDF weighted unique heat vector. Therefore, when the target question and the candidate question contain synonyms, the similarity calculation effect of this traditional method is not good. In order to solve this problem, this paper proposes an improved text similarity algorithm based on synonym word forest. This algorithm replaces the existing synonyms in the text in advance based on the synonym word forest based on the traditional method, thereby reducing or eliminating the effect of synonyms on the calculation of text similarity. The improved text similarity algorithm based on synonymous word forest has the following steps:

Step1. Synonym word forest is generated based on the “Synonym Cilin Expansion Version of Harbin Institute of Technology Social Computing and Information Retrieval Research Center” [14]. For ease of processing, only the word groups that overlap with the content of the corpus are retained;

Step2. Based on the synonym word forest obtained in the previous step, replace the vocabulary with synonyms in the target problem and candidate problem, and replace it with the first word of the corresponding word group in the synonym word forest;

Step 3. Calculate the TF-IDF weighted one-hot vectors of the target problem and candidate problem after replacing synonyms, and calculate the cosine similarity to obtain the text similarity between the target problem and the candidate problem.

Since the automatic question answering method proposed in this paper is designed for a small-scale corpus, the number of synonymous word groups involved is usually small, so the time and space complexity of the improved text similarity algorithm based on the synonym word forest is higher than before. Based on the question-and-answer corpus provided by a hotel, this paper tests the effect of the improved text similarity algorithm. Some test results are shown in table 1.

| Sentence A                          | Sentence B                          | Before | After |
|-------------------------------------|-------------------------------------|--------|-------|
| What’s the hotel’s special features? | What are the characteristics of the hotel? | 0.248  | 1.0   |
| Where is Japanese cuisine at the hotel? | Where Japanese food is available at the hotel? | 0.109  | 1.0   |
| Can the hotel wash clothes?         | Does the hotel have a laundry service? | 0.163  | 1.0   |
| Does the room rate include breakfast? | Does the room rate include breakfast? | 0.517  | 1.0   |
| When does the park open?            | What time is open in the park?       | 0.323  | 1.0   |

It can be seen from table 1 that the improved text similarity algorithm can effectively solve the problems caused by the phenomenon of synonymous different words to the traditional text similarity calculation, thereby obtaining more accurate text similarity information.

3.3. Intent Recognition Method Based on Slot Filling

The method based on text similarity calculation can only realize a single round of dialogue. In order to realize the support of multiple rounds of dialogue, this paper designs a slot-based intent recognition method to perform task-oriented multiple rounds of dialogue.

Before introducing intention recognition, it is necessary to clarify and analyze the concept of multiple rounds of dialogue. When the information in the target problem is sufficient and the precise semantics can be obtained directly, the problem can be handled according to a single round of dialogue strategy; when the target problem is missing, omitted, or implies some important information,
it is necessary to analyze its intention, and explicitly supplement the missing information and get the precise semantics of the target problem. The multi-round dialogue strategy is designed for this situation.

Studies have shown that in English listening practice, it is possible to make reasonable guesses by capturing key information such as feature words in sentences [15]. In daily conversations, people’s understanding of semantics is often based only on some of the key words. Therefore, this paper realizes the intention recognition based on the feature word matching method.

This article uses XML language to store the dialog system intent recognition template, as shown in figure 2. In this template, the <topic> tag corresponds to the intent topic, and the <keywords> tag. And <keyword> tags are the feature words that should satisfy the relationship of “and” or “or”. The <slot> tag is the intention data slot. Each <element> corresponds to a condition element and is defined in the <ques> tag questioning statement at this condition.

![Figure 2. Intention recognition template.](image)

According to the intent recognition template, it is easy to classify the intent of the target problem, but in order to accurately locate the intent of the target problem, it is necessary to supplement the conditional elements omitted in the problem. The supplement of the omitted elements is achieved through the headword extraction and questioning. The specific steps are as follows:

Step1. For all the input target questions, based on the pre-defined information in the part-of-speech and participle lexicon, extract the headwords that may be omitted in the question below and save;

Step2. After the intent recognition classification is successful, if there is a lack of conditional elements, the headword extracted above is first filled in to match the candidate question. If the matching is successful, the intent recognition is successful;

Step3. If the matching fails after filling in the head word above, the missing condition elements are questioned according to the content of the <ques> tag in the intent recognition template.

4. Experiment

The use of sections to divide the text of the paper is optional and left as a decision for the author. Where the author wishes to divide the paper into sections the formatting shown in table 2 should be used.

4.1. Experimental Environment

The automatic question-and-answer method proposed in this paper is a non-statistical algorithm, the implementation of the algorithm requires lower requirements for the system hardware and software environment, can run on ordinary PCs. We use a PC with Intel(R) Core(TM) i5-3230M CPU @ 2.60GHz, and memory of 8.0GB. Our operation system is Windows 10 and programming language is Python 3.7.3. The data encoding is UTF-8.
4.2. Analysis of Experimental Data and Result

The main goal of the experiment is to test the question of the automatic question-and-answer method proposed in this paper, and to see whether the accuracy and response time are reasonable. To eliminate the effect of unrelated domain issues on results, this paper uses 10 volunteers to conduct random question sand data with the system under conditions in a given problem area as experimental data.

10 volunteers and the system had 100 rounds of random conversations, and 921 valid target questions were removed from the overrange questions. The experimental data for these 921 valid target questions was tested at different values at the thresholds A and B, and the q and answer accuracy and average response time were shown in table 2.

The question domain of our automatic question answering system includes:
- Hotel conditions—overview, services, room status, charges, surrounding;
- Hotel interior—location, characteristics, business hours, capacity;
- Information of the city—overview, train station related, bus station related, airport related;
- Sights—overview, location, distance, business hours, charging situation, transportation.

As we can see, in this experiment, when A and B are valued at 0.9 and 0.8 respectively, the accuracy of this automated question-answer method can reach 92.9%. The average response time at this point is 38.1ms, which meets the general requirements for a small-scale automated question-answer system.

Table 2. Accuracy and average response time with different values of A and B.

| Threshold A | Threshold B | Question and answer accuracy | Average response time /ms |
|-------------|-------------|------------------------------|--------------------------|
| 0.95        | 0.9         | 91.7%                        | 39.4                     |
| 0.95        | 0.85        | 92.3%                        | 37.9                     |
| 0.95        | 0.8         | 92.9%                        | 38.9                     |
| 0.9         | 0.85        | 92.3%                        | 39.3                     |
| 0.9         | 0.8         | 92.9%                        | 38.1                     |
| 0.9         | 0.75        | 92.0%                        | 36.8                     |
| 0.85        | 0.8         | 92.9%                        | 38.4                     |
| 0.85        | 0.75        | 92.0%                        | 37.7                     |
| 0.85        | 0.7         | 91.1%                        | 37.0                     |

5. Conclusion

Aiming at the problem of lack of language and materials often encountered in the development of automatic question and answer system, this paper puts forward an automatic question-and-answer method for small-scale language materials, which first improves the traditional TF-IDF-based text similarity algorithm, and combines it with the intent recognition method to improve the accuracy of single-round dialogue and increase the support for multi-round dialogue. Experimental tests on an automated question-and-answer system based on a hotel’s language data show edified that the automated question-and-answer method can also achieve high q and a-answer accuracy when the language is small. The intent identification template required for this method still needs to be written manually, and the automatic or semi-automatic generation method of the intent identification library will continue in the next step.

References
[1] Turing A M 1950 Computing machinery and intelligence Mind 59 (236) 433-460.
[2] 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.
[3] A.L.I.C.E. (n.d.) http://www.alicebot.org.
[4] Wang H and Li B 2018 Research progress of chatbot system Computer Applications and Software 35 (12) 1-6.
[5] Feng D 2013 Research and Implementation of the Graduate Admissions Counselling Intelligent Chat Robot Based on ALICE (Chengdu: Southwest Jiaotong University) Doctoral Dissertation.

[6] Talk-Bot Page. (n.d.). http://www.frontier.net.net/~wcowart/index.shtml.

[7] ChatScript download. (n.d.). http://source forge.net/projects/chatscript/.

[8] Cho K, Van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H and Bengio Y 2014 Learning phrase representations using RNN encoder-decoder for statistical machine translation Empirical Methods in Natural Language Processing 1724-1734.

[9] Tai K S, Socher R and Manning C D 2015 Improved semantic representations from tree-structured long short-term memory networks International Joint Conference on Natural Language Processing pp 1556-1566.

[10] Zhou B, Sun C, Lin L, et al. 2018 LSTM based question answering for large scale knowledge base Acta Scientiarum Naturalium Universitatis Pekinensis 54 (2) 286-292.

[11] Feng M, Xiang B, Glass M R, Wang L and Zhou B 2015 Applying deep learning to answer selection: A study and an open task IEEE Automatic Speech Recognition and Understanding Workshop pp 813-820.

[12] Li L and Li H 2018 Computing text similarity based on concept vector space Data Analysis and Knowledge Discovery 2 (5) 48-58.

[13] Zhou Y, Li J and Cai S 2019 A semantic similarity method of sentences based on Tongyici Cilin and its application in question answering system Computer Applications and Software 36 (8) 65-68.

[14] LTP-Cloud (n.d.) http://www. ltp-cloud.com/download.

[15] Cheng X 2018 A Study on Prediction Strategy in Junior English Listening Teaching Based on Figure-Ground Theory (Chengdu: Sichuan Normal University) Doctoral Dissertation.