Research on Question and Answer System for Open Domain

Fangtao Yang[1, 2, a], Fucheng Wan[1, 2, b, *], Ning Ma[1,2, c], Tiantian Wu[1, 2, d]

[1]Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education Northwest Minzu University, Lanzhou, Gansu 730000, China

[2]Key Laboratory of China's Ethnic Languages and Intelligent Processing of Gansu Province, North west Minzu University, Lanzhou, Gansu 730000, China

[a]780873497@qq.com, [b, *] 306261663@qq.com, [c] 6105112@qq.com, [d]1316170316@qq.com

Abstract. With the rapid development of computer technology and the Internet, people expect to get accurate target information from massive information efficiently. Compared with the traditional search engine based on keywords, the question-and-answer system can better meet people's retrieval needs. As an advanced form of information retrieval, q & a system allows Chinese natural language to be used as the query condition, and directly returns the results to users in the form of answers, which greatly improves users' retrieval satisfaction and time cost, Question answering system generally consists of three parts: question analysis, information retrieval and answer extraction.

Keywords: question answering, question analysis, information retrieval, answer extraction.

1. Introduction

The automatic question answering system is a system that automatically answers the questions raised by the user to the user's desired answers through a computer. Because now is the information age, in order to use the resources on the network to find the information you need, we have the traditional search engine Google, Sohu, Yahoo, etc. We only need to input some keywords, the engine can find Out of the relevant page. But these search engines also have a lot of improvements. For example, some engines find too many web pages, making it difficult for users to quickly find the results they want, which may cause users to waste more time. In addition, these traditional Search engines cannot understand the natural language that users enter. Sometimes users can't clearly enter their exact intent, and search engines can't find the answer they want. Unlike existing search engines, the Q&A system is an advanced form of information service that does not return a list of documents based on keyword matching, but instead returns a clear natural language answer. The question-and-answer system is integrated with natural language processing technology. By understanding the problem, it directly provides users with the results they want. This is like an erudite scholar who can answer many questions as quickly as possible. For example, the user input "Where is the Northwest University for Nationalities?" The question and answer system can directly give the result "Northwest University for Nationalities in Lanzhou". Over the past few years, due to the rapid advancement of artificial intelligence, the automatic question and answer has become the focus of attention and has broad prospects for development.

2. Research Overview

In recent years, search engines based on keyword search have developed rapidly, such as Baidu, Sogou, Yahoo, etc., providing better technical support for users to quickly search for information from the Internet. However, these search engines are not able to effectively indicate people's complex search intent and it is difficult to get results that meet the requirements. In order to solve this problem, professional technicians are also trying to study a more efficient and user-friendly search engine technology - question answering system. At the same time, many search engine companies have launched their question and answer system, such as Baidu know, Sogou, and Tao, and so on, but also got a lot of attention. According to the field of the search, the question and answer system is divided into a restricted domain question answering system and an open domain question answering system as well as a frequently asked question set question answering system. The restricted domain question

answering system is designed for specific areas and needs to be answered with domain knowledge. With the advancement of the times, users are no longer limited to a specific field, interest and scope are continuing to increase, users' query methods are more flexible, and the complexity of questions is increasing. It is no longer just a few simple questions, but it also needs to search for relatively difficult questions, such as: "How to learn natural dialectics", "How to learn linear algebra" and so on, like this is shown in Fig.1

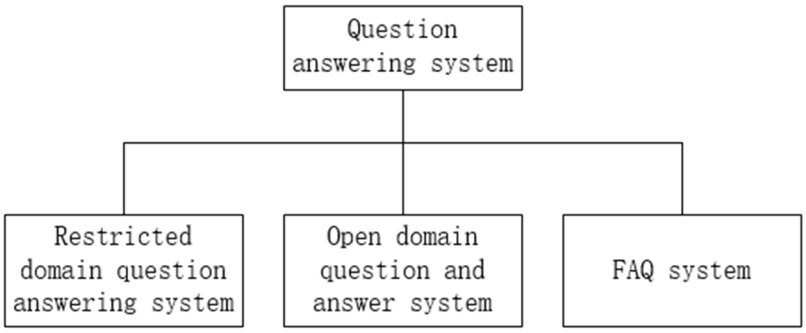

Figure 1. Question and answer system are classified according to problem areas

Although the concept of the open domain question answering system has not been proposed for a long time, some well-established systems have been formed. The first online Q&A system START was developed by MIT Bons Katz et al. Since its release in December 1993, it has answered countless questions, including location, film, humanities, technology, geography, climate, etc. . Ask Jeeves is also a good open domain question answering system. Compared to START, it returns not the exact result, but the relevant fragment containing the result. Brain boost and Answer Bus are also popular online systems that return statements with answers.

In order to promote the progress and improvement of the open domain question and answer system, the well-known text search conference (TREC) in 1999 organized a back-to-back question and answer contest for the question and answer system, and many research institutions, companies, and universities also participated. In October 2000, the Open Domain Question Answering System was established in the 38th International Conference on Computational Linguistics held in Hong Kong. Since the TREC and ACL meetings have continuously organized QA assessments, the open domain Q&A system has quickly become a hot spot for exploration. In recent years, the number of scientific research institutions that have been researching and answering questions in China has gradually increased. Harbin Institute of Technology, Peking University, Tsinghua University, and Chinese Academy of Sciences have also done a lot of research, and many organizations have achieved good results in the recent TREC QA Track evaluation. The response.

3. System Framework

The automatic question answering system studied in this paper consists of three modules: question analysis module, information retrieval module and answer extraction module. The process is as follows: First, analyze the questions raised by the user, including word segmentation, extracting sentence keywords, keyword expansion, question classification, etc. Second, search for keywords, then search for answers, and process the results. Third, extract and submit the results by similarity calculation and sorting. As shown in Figure 2:

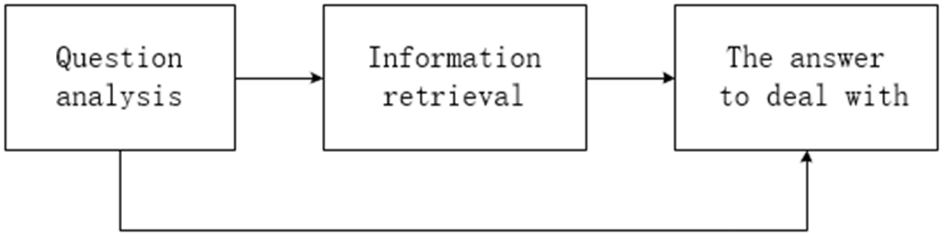

Figure 2. Three major processing modules of the Q&A system

4. Question Analysis Module

Question analysis is the first step of the system analysis. This step plays an important role in the subsequent series of processing. The question analysis module is very important in the question and answer system, and plays a decisive role for the other two modules, and this part is also the main task in the question and answer system. Nowadays, the performance of all aspects of the automatic question and answer is still insufficient. The first major factor is that the system cannot clearly understand the questions raised by users. Question analysis is the application of syntactic structure analysis, dependency analysis, block analysis and semantic analysis methods such as semantic role labeling and other analytical techniques to process questions to guide the choice of answers. Problem understanding is a difficult part, and it is the ultimate ideal effect that the question analysis wants to achieve.

The first step in the Q&A system is to accept the questions that the user enters in a natural language, and to analyze and deal with them. The best result is that the system can fully understand and give a clear answer. This is the Q&A system. The main purpose.

4.1 Question Classfication

Question classification Popular point refers to the classification of questions by the type of answer corresponding to the question, which is the main role of question analysis, because the answer type will affect the next answer extraction module. For example, the location extraction module will initially consider the location that appears in the document as an answer candidate set. Defining the type of question according to the question word is the most common and simple classification. In English, the classic has 5W1H (What, Who, When, Where, Why, How). Similarly, there is a similar "what" in Chinese. , "how" and other types of question words. However, the granularity of this method is not very detailed, especially for where and how to ask questions, and the type of answer corresponding to it is also very diverse. As shown in Table 1:

Table 1. Common problem types

question type	Question words	example
Inquirer	Who / who	Who discovered the gravitation?
Inquiry time	When / when / that year...	What year is it now?
Number of inquiries	How many / how big / how high...	How tall is Xiao Ming?
Inquiry definition	What is / what?	What is carbon dioxide?
Ask for a place or location	Where / where / what place	Where is the Forbidden City?
Ask why	why	Why is it dark?
other	-	-

4.2 Keyword Extraction

Question analysis also has an important feature, keyword extraction. In the next information retrieval module, we have to choose to query some of the keywords, and the query needs to be adjusted when necessary. But no matter how it is adjusted, it must include the subject of the problem. Usually, we can parse the syntax of the question to get the keyword of the question, then select the

keyword and the modifier related to the keyword as the subject. How to choose the appropriate center word size is the main problem to be solved. Cui initiated a method based on the selection of external resource phrases, and submitted the keywords in the question to the search engine. From the answers returned by the search engine, the mutual information of the various words was found. Only the keyword combination above a certain level can be used. It is called a phrase, so the sequence of words forms the subject of the question. Query keywords are also included in the analysis of problems in some systems. Keywords are mainly composed of nouns, verbs, adjectives, and qualified adverbs. Some question-and-answer systems also classify keywords into two categories: general keywords and "must contain" keywords. The so-called "must contain" keyword means that these keywords must be included in the response sentence, and the general keyword cannot be included in the response sentence. As shown in Table 2:

Table 2. Part of the question words

what	Who	What	Which	Which	How to do
how	what	What to do	how about it	many	He Yue
where	Why	why	where	How many	how is it
a few	which year	Why	First few	How	Why
How high	What happened?	no	What?	What time	Ye

4.3 Keyword Expansion

In order to improve the retrieval rate of the system, the question and answer system usually needs to expand the keywords. In the response sentence, some words may not be the key words of the question, but the synonymous extension of those keywords. Although the keyword expansion improves the retrieval rate of the system, the inappropriate expansion may greatly reduce the accuracy, so most question and answer systems expand the keywords very carefully. Therefore, most question and answer systems add a lot of restrictions to the expansion of keywords. For example, keyword expansion can only be used for nouns, and can be extended by Word net or other synonym dictionary. Some question-and-answer systems use statistical methods to expand keywords, but this method requires a large number of Q&A corpora for training. No matter what kind of question the answers generated have some common features. Moreover, some question and answer systems also expand the keywords by retrieving related documents returned. However, the accuracy of the expanded keywords is not so high, so the accuracy of the system is expected to improve. Many question and answer systems add weights to the keywords to distinguish their importance.

5. Information Retrieval Module

Different from the traditional information retrieval, the information retrieval module of the automatic question answering system searches the relevant documents with the previously extracted keywords, and finally returns the documents most relevant to the answers. The quality of the information retrieval results is related to the accuracy of the results of the question and answer system. Even for well-known search sites like Baidu and Google, the results of information retrieval can sometimes be disappointing, because the returned documents are sometimes irrelevant to the questions asked by users, and sometimes even the correct answers will be Being discharged to the back, it is not able to meet the needs of users well. These situations are essentially due to problems with indexing and similarity calculations that are not well resolved.

In the information retrieval module, the first is to index the document library, which is important for finding keywords related documents in a large number of documents. It extracts the results first, then prepossess the results, and then calculates the correlation. By continuously improving the information retrieval module, the returned paragraphs can be related paragraphs, or even sentences, which can greatly improve the overall performance of the question and answer system.

The information retrieval structure is shown in Figure 3 below:

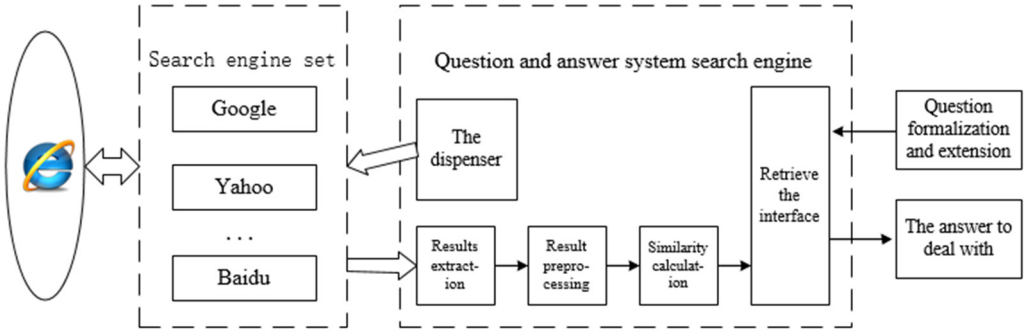

Figure 3. Information Retrieval Structure

Information retrieval has always been a research hot spot of traditional information retrieval and automatic question and answer. Most of the previous question and answer systems were based on density statistics. How to use various natural language information more effectively and adopt more efficient machine learning algorithms in the sorting or reordering of text segments is an important research direction of information retrieval in the future.

6. Answer Extraction Module

The answer extraction module uses the relevant analysis and reasoning mechanism to retrieve the answers that the user wants from a large number of documents and massive web page information. When extracting the answers, the method used is related to the whole question and answer system. Performance, so the method of selecting the answer extraction is very important. The answer extraction module is generally divided into two parts: one is to generate a candidate answer set; the other is to use different methods to extract the answer.

The answer extraction is generally divided into the following steps:

Using the information retrieval module to analyze the paragraphs or sentences from a large number of documents and massive web page information, and then divide them into individual sentences and use them as candidate sets of answers;

Through the question analysis, the type of the problem is obtained, and the candidate answer set is further processed to reduce the sentence irrelevant to the answer of the question, thereby improving the accuracy of the answer;

Calculate the similarity between the question and the answer;

The candidate answers are sorted according to the level of similarity calculation results, different weights are assigned to them, and sentences with high similarity are returned.

The work flow diagram of the answer source is shown in the following figure 4:

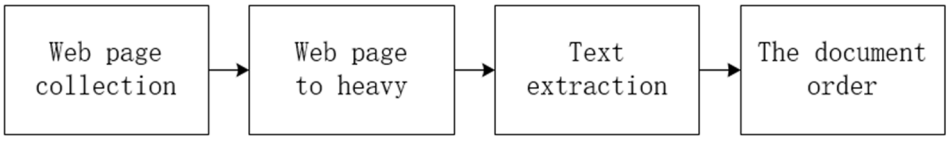

Figure 4. Schematic diagram of the work flow for the answer source acquisition

After getting the set of candidate answers, there are four main ways to get the best answer to the question:

The answer is extracted by pattern matching: according to different problems, the corresponding mode is formulated. The general mode is obtained by machine learning or manual method. Due to the complexity of the problem, it is often difficult to cover and expand by manual methods, so machine learning is often used. Methods.

The statistical model is used to extract the answer: firstly, the retrieved sentences are analyzed, the question words in the sentences are retained, then the candidate sentences containing the answers are searched and the sentences are analyzed, and finally the statistical translation model is used to extract the answers. This method mainly relies on the scale and quality of the training corpus, and it is difficult to obtain large-scale training corpus for the open domain question answering system.

The answer is extracted based on information retrieval: the algorithm mainly searches by keywords, and generally divides keywords into common keywords, extended keywords, basic noun phrases, quotation words and other keywords. This method is relatively simple and relatively easy to implement, considering only discrete words, regardless of the relationship between words and words.

The answer is based on natural language understanding: mainly through a certain degree of syntactic analysis and semantic analysis through the question and answer text, and then achieve reasoning. Because the technology of natural language processing is still not mature enough, the deeper technology has not yet reached a practical level.

7. Conclusion

With the exponential growth of network information, people put forward higher requirements for the effectiveness of search engines and the ease of use, because traditional search engines often compare irrelevant information that is often returned when a user asks a specific problem. Many, can not provide the user really want the answer, although many large search engine companies, such as Baidu, Google, etc. are constantly improving the performance of search engines, but the traditional search engine still has certain limitations. With the advent of the automatic question answering system, the effect of the search has been greatly improved. The question answering system is, in a certain sense, a new generation of search engines that integrates knowledge representation, information retrieval, natural language processing and intelligent reasoning. Although the overall performance of the question-and-answer system is much higher than that of the traditional search engine, the question-and-answer system does not answer any questions as accurately as humans, and then proposes an open-domain question-and-answer system that is no longer specific. The field, and mainly the Internet-oriented question and answer system.

The open-domain question-and-answer system is more difficult because of its wide range of processing objects and complex content. Although the accuracy of the evaluation results is relatively low, the Internet is the most frequently contacted by people every day. Therefore, it is open to the domain. The question and answer system has become more and more the focus and hotspot of research. It is believed that with the promotion of its broad application prospects, in the near future, the system of question and answer will have a breakthrough, and there will be a qualitative leap to meet the needs of users.

Acknowledgments

This research is supported by The National Natural Science Fund (NO. 61762076).

References

[1]. Tang Zhaoxia; answer extraction algorithm for Chinese question answering system with multi-feature fusion[j];Journal of Guizhou University(Natural Science Edition);2011-05.

[2]. Wu Youzheng, Zhao Jun, Duan Xiangyu, Xu Bo; A review of question-and-answer search technology and evaluation research [j]; Chinese Journal of Information; 2005-03.

[3]. Zheng Shifu, Liu Ting, Qin Bing, Li Sheng; Overview of Automatic Question and Answer [j]; Chinese Journal of Information; 2002 06.

[4]. Huang Bo; Research and implementation of answer extraction in Chinese question answering system [d]; Jilin University; 2010.

[5]. Luo Lijun; Research and implementation of several technologies in Chinese information processing [d]; Liaoning University of Science and Technology; 2008.

[6]. Wang Zhenghua; Han Yongguo;Design and implementation of automatic question answering system[j]; Software Guide; 2014 09.

[7]. Cai Gangshan;Chinese automatic question answering system research[d]; Huazhong University of Science and Technology; 2007.

[8]. Hu Dawei;Research and Implementation of Answer Acquisition Method for Question Answering System[d]; University of Science and Technology of China; 2008.

[9]. Zhang Hua, li chao. Design and implementation of intelligent question-and-answer system for Java courses [J]. Computer era,2018(12):12-15.

[10]. Zhang Yue, Muyun Yang, Dequan Zheng, Information retrieval automatic evaluation method for question and answer system [J]. Intelligent computer and application,2019,9(02):262-268.

[11]. Wenzheng Feng, Jie Tang. Answer selection model of fusion depth matching features [J]. Chinese journal of information technology,2019,33(01):118-124.

[12]. Xu Xiong. Research on q&a system based on deep learning [J]. Journal of hubei normal university (natural science edition),2019,39(01):10-18.

[13]. Wang Feihong. Design of intelligent question-answering system based on automatically generated knowledge base [J]. China science and technology information,2018(12):50-52.

[14]. Chou lei. Research and implementation of key technologies of internet-based automatic question answering system [D]. Northwest university,2018.