Research on Answer Extraction for Automatic Question Answering System

Tiantian Wu[1, a], Hongzhi Yu[1, b], Fucheng Wan[2, c, *] and Fangtao Yang[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, Northwest Minzu University, Lanzhou, Gansu 730000, China

[a]1316170316@qq.com, [b]3095911462@qq.com, [c,*] 306261663@qq.com, [d]780873497@qq.com

Abstract. The question answering system as a hot issue in the field of NLP study, and answer extraction as the core part of the automatic question answering system, its effect is good or bad is directly related to the performance of question answering system, this paper introduces the technical process of answer extraction module and the use of pattern matching technology, gradually shift from the traditional pattern matching technology to the machine learning method of pattern matching technology, and also calculate the questions and answers in the answer extraction module are several methods of similarity. Since the performance of the automatic question answering system depends on the algorithm used for answer extraction to a large extent, the selection of the algorithm for answer extraction is also crucial, and the accuracy of the answer returned to the user can be improved by improving the algorithm.

Keywords: Automatic Question Answering, Answer Extraction, Pattern Matching, Similarity.

1. Introduction

With the rapid development of Internet, appear in front of people is increasingly huge amounts of information, users expect from huge amounts of information quickly find what they need in the desire of the target information more and more strong, search engine providers in recent years more and more attention to improve the user experience, and spent a great deal of manpower and material resources and financial resources to further improve the function of search engine, but still can't meet the needs of users, has certain limitations, such as search engines too much useless information feedback to customers, a waste of resources and time, at the same time, it will produce certain misleading to customers, search can not find the information users really need, can not well meet the needs of users, so people put forward higher requirements for the effect of search engine and the convenience of use, with the emergence of automatic question and answer system, greatly improve the effect of search. Answer extraction module is the core part of automatic question answering system, so the performance of question answering system largely depends on the effect of answer extraction module.

2. Research Overview

Back in the 1960s, when artificial intelligence was still in its infancy, it was proposed that computers should understand the questions people asked in natural language and then return the answers to their questions, which is referred to as the question answering system now. The question-and-answer system contains not only natural language processing technology, but also information retrieval technology, which is equivalent to the new generation of search engines. In 1993, MIT's artificial intelligence lab developed the world's first WQA system -- START, which can answer many English questions related to culture, history, geography, technology and entertainment, and the number of questions can be quite large, up to millions. It is a comprehensive utilization of WQA and KBQA related technology of hybrid question answering system, for the user, it will first search the knowledge base, when the user when the proposed problem can be found in the knowledge base, is the repository directly to the corresponding answer returned to the user, only when the answer does

not exist in the knowledge base to meet the needs of users, can use search engines to look up the information correlated with question, and then the extract out the answer returned to the user[1].

Answer extraction is to extract the pattern of the problem first, and then classify the problem through pattern matching, and then get the category of the problem. In recent years, there has been a gradual shift from the manual organization approach to the machine learning approach. In the early 21st century, Ravichandran proposed a method of using machine learning, that is, using supervised learning method to train the answers provided by the user < question, answer> as the training expectation, and then conduct Web search. In 2004, Du et al. proposed a supervised learning method similar to Ravichandran, and the difference between the two is reflected in the classification and mode representation of questions[2]. For supervised learning this kind of machine algorithm, because the quality of the algorithm is largely dependent on the training provided by the user, that is, < question, answer > pairs. Although its for certain types of questions has good performance, but due to the representation of the answers are varied, complicated, which has certain diversity and complexity, are unlikely to let the user to provide cover all questions of question and answer mode, the algorithm of q&a mode will cause a certain degree of influence. For example, users asked questions like "where was lu xun's birthplace? " When such kind of question, its answer may be "zhejiang", "zhejiang province", "zhejiang shaoxing" and "zhejiang shaoxing government office kuaiji county" and so on.

Wu et al. put forward another machine learning method, that is, unsupervised learning method is adopted to extract the answer mode. First, the answer mode is extracted from the Internet, and then it is applied to the answer mode of Chinese question answering system[3]. This method is different from Ravichandran, Du the supervised learning machine method, put forward the method without user to provide answers to as expected of training, only requires the user to give examples of each question type, algorithm can through the Web retrieval, subject classification, pattern extraction, vertical clustering and horizontal clustering and other steps to complete the types of questions the answer mode of learning.

Tan M applies a general deep learning framework to complete the answer selection task, which does not rely on manually defined features or language tools. Its basic framework is to build the embedding vector of questions and answers based on the biLSTM model, and measure their tightness through cosine similarity[4]. One direction is to combine the convolutional neural network with the basic framework to define a more complex question-and-answer representation. The other direction is to use a simple and effective attentional mechanism to generate answers based on the context of the question.

Wang D proposed a sentence selection method based on the combination of multi-layer two-way LSTM model and keyword matching[5]. It demonstrates the knowledge of distributed and symbolic representations of complementary types of coding, both of which contribute to the identification of answer sentences, the most important of which do not require any syntactic features or external resources.

Wang B makes qualitative and quantitative analysis of traditional model-based attention deficit.On this basis, three new RNN models were proposed, in which attention information was added before the hidden representation of RNN, which showed certain advantages in the representation of sentences and made new technical achievements in the task of selecting answers[6].

Shen D discussed the correlation of the dependent path of candidate answer sorting. By using the correlation measure, the dependence of the candidate answers is compared, and the question phrases in the sentence are mapped to the corresponding relationships of the questions[7]. Different from previous studies, an approximate phrase mapping algorithm is proposed, and the mapping score is included in the correlation measure. These correlations are further incorporated into a Maximum Entropy based ranking model, which estimates the path weight through training. Experimental results show that this method is significantly superior to the existing grammar-based method in MRR, up to 20%.

3. Answer Extraction

Answer extraction is to use the relevant documents returned from the information retrieval module, do some pre-processing on the relevant documents, extract the answers to the questions, generate a candidate answer set, and then use some answer selection algorithm to sort the candidate answers, return the answers with the highest weight, and finally output the answers. The answer extraction module is generally composed of candidate answer extraction and candidate answer sorting. The common candidate answer extraction method is named entity method. When the candidate answer is sorted, it is most important to calculate the similarity between the question sentence and the candidate answer sentence.

Question analysis module, information retrieval module and answer extraction module are three essential parts of an automatic question answering system.

(1)Problem analysis module: the problem analysis module is responsible for the analysis and processing of the problems raised by users, including determining the type of the problem, extracting the keywords of the problem, extending the keywords appropriately and semantic analysis of the questions.

(2)Information retrieval module: the information retrieval module mainly USES the query keywords obtained from the question analysis module, and then retrieves some information related to the questions raised by users based on certain retrieval methods, and then returns relevant documents to the answer extraction module[8].

(3)Answer extraction module: according to the documents returned from the information retrieval module, the answer extraction module extracts entities with the same type of question answer by using relevant analysis and reasoning mechanism, and then adopts a strategy to calculate the weight of each candidate answer and return the answer with the highest weight to the user[9].

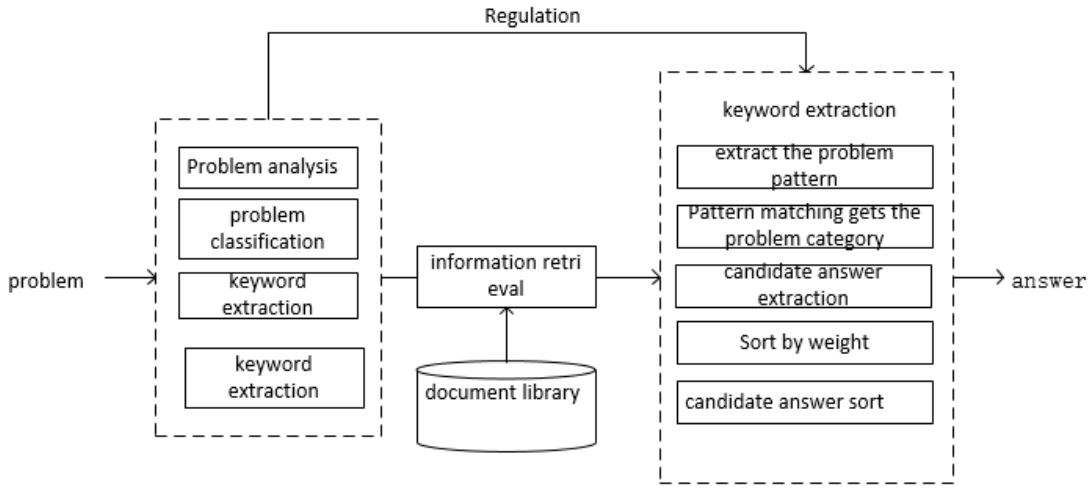

Figure 1. Question answering system structure

3.1 The Form of the Answer

Answer extraction shall be based on the type of question, which can be divided into the following four types:

(1)Use sentences as the answer: firstly, divide the documents retrieved from the information retrieval module into sentences, calculate the weight of each sentence according to some algorithm, and sort them in order according to the weight of the sentence, and then reorder the candidate answers according to the type of the question.

(2)With words or phrases as the answer (such as asking the name of time, place and organization) : this type of answer is mainly for the fact that a large number of relatively short answers exist in the question answering system, in which nouns and named entities are the main ones.

(3)Take the article as the answer: on the one hand, sometimes users not only want to know the specific answer to the question, but also want to have a deep understanding and analysis of the article in which the answer is found; On the other hand, because the whole article describes the answer to the question, the system cannot locate the specific answer, so it will return the whole article as the answer.

3.2 Determination of Candidate Answers

Firstly, the document returned by the information retrieval module is preprocessed and broken into individual sentences. Then word segmentation and part-of-speech tagging are carried out. The result of question analysis is used to exclude the sentences that do not contain the expected answer type to get the candidate answer.

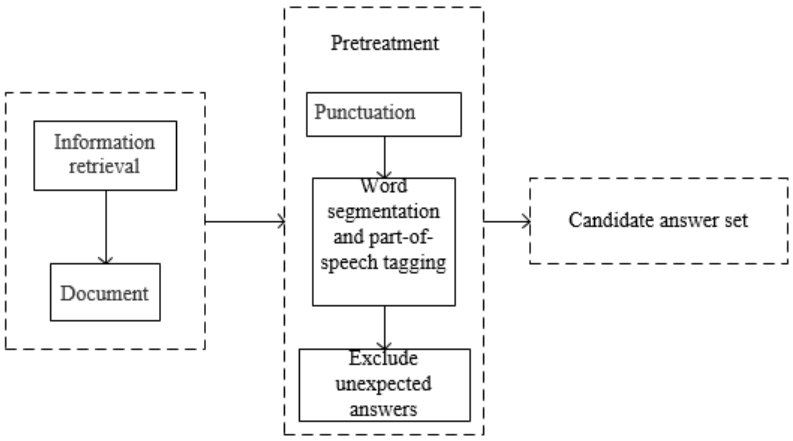

Figure 2. Candidate answer determination

3.3 Sorting of Candidate Answer

When extracting candidate answers, the ranking of candidate answers depends on certain strategies and algorithms. The similarity between the calculated question and the answer is to improve the ranking effect of the candidate answers, so that the answer returned to the user is closer to the target answer and better meet the needs of the user.

Depth according to the analysis of the statement, the similarity calculation method mainly exists in three ways: one is the method based on vector space model, the method is wrong sentence grammatical structure analysis, the corresponding statement sentence similarity measure mechanism can only use the surface layer of information, of the word of a sentence of word frequency, such as part-of-speech information. Because it is a statistical method, only when the sentence contains more words, the related words will repeat, the effect of this statistical method can be reflected; The second is a complete syntactic and semantic analysis of the sentence, which is a kind of deep structural analysis method. First, we need to find out the dependency relationship, and calculate the similarity on the basis of the dependency analysis results. For example, li et al. integrated Chinese dependency syntactic information into question sentence analysis[2], while Oliva et al. integrated syntactic role information into question sentence analysis model[10]. Thirdly, fusion method. Xiong et al. proposed a lda-based similarity calculation method for questions, which integrated the statistical information, semantic information and subject information of questions to calculate the similarity of questions[11]. Saric et al. adopted the method of word overlapping and syntactic combination[12]. Although these methods have achieved certain results, they are still affected by the performance of the question answering system and cannot maintain high computational accuracy.

4. Evaluation

The TREC QA evaluation funded by the National Institute of Standards and Technology (NIST) is the most influential in the research of automatic question answering system. Its evaluation tasks

and indicators change from year to year. The main types of questions include factual questions, phenotypic questions, definitional questions, contextual questions, paragraph questions and other types of questions.

TREC QA evaluation indexes mainly include: mean reciprocal rank (MRR), accuracy, etc. The calculation formula of MRR is as follows:

$$\mathrm{MRR} = \frac{\sum_{i=1}^{N} \frac{1}{\text{The position of the standard answer in the sort results given in the system}}}{N} \tag{1}$$

The calculation formula of accuracy is as follows:

$$\text{Accuracy} = \frac{\text{Number of questions answered correctly}}{\text{Total number of questions}} \tag{2}$$

5. Conclusion

From the research and analysis of the question answering system, the answer extraction module contains two parts: candidate answer extraction and candidate answer sorting. In order to improve the accuracy of the answer, certain strategies must be adopted to improve the two parts. We should combine the shallow grammar analysis with the deep semantic analysis, and pay attention to the combination of machine learning and deep learning methods, so as to provide users with more satisfactory answers.

But in the process of candidate information retrieval and answer extraction problems more noise, syntactic structure is not complete, when calculating the similarity of questions and answers with poor scalability problems, etc., some technology may not be very mature, further in-depth study in the future, efforts to improve, to improve the accuracy of automatic question answering system, believe in the near future answer extraction accuracy will be greatly improved.

Acknowledgements

The authors would like to express their appreciation for the financial support received from National Natural Science Fund (NO.61762076) and Northwest Minzu University for permission to publish this paper.

References

[1]. Li Zhoujun, Li Shuihua. Survey on Web-based Question Answering[J]. Computer Science, 2017, 44(06):1-7+42.

[2]. Li Bin, Liu Ting, Qin Bing, Li Sheng. Chinese Sentence Similarity Computing Based on Semantic Dependency Relationship Analysis[J]. Application Research of Computers. 2003 (12) :15-17.

[3]. Wu Youzheng, Zhao Jun, Xu Bo. Unsupervised Answer Pattern Acquisition[J]. Journal of Chinese Information Processing,2007, 21(2):69-76.

[4]. Tan M, Santos C, Xiang B, et al. Lstm-based deep learning models for non-factoid answer selection[J]. arXiv preprint arXiv:1511.04108, 2015.

[5]. Wang D, Nyberg E. A long short-term memory model for answer sentence selection in question answering[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2015, 2: 707-712.

[6]. Wang B, Liu K, Zhao J. Inner attention based recurrent neural networks for answer selection[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016, 1: 1288-1297.

[7]. Shen D, Klakow D. Exploring correlation of dependency relation paths for answer extraction[C]// International Conference on Computational Linguistics & the Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2006.

[8]. Zheng Shifu, Liu Ting, Qin Bing, Li Sheng. Overview of Question-Answering[J]. Journal of Chinese Information Processing,2002, (06):46-52.

[9]. Zong Chengqing, Statistical Natural Language Processing[M]. Bei Jing: Tsinghua University Press, 2013:450-451.

[10]. Oliva J, Serrano J I, Castillo M D D, et al. SyMSS: A syntax-based measure for short-text semantic similarity[J]. Data & Knowledge Engineering, 2011, 70(4):390-405.

[11]. Xiong Daping, Wang Jian, Lin Hongfei. An LDA-based Approach to Finding Similar Questions for Community Question Answer[J]. Journal of Chinese Information Processing, 2012, 26(5):40-46.

[12]. Frane Šarić, Goran Glavaš, Karan M, et al. TakeLab: Systems for Measuring Semantic Text Similarity[J]. In Proceedings of SemEval-2012, 2012.