Application Research of a Practical and New Intelligent Question Answering System

Hua Zhou*, Yaofei Tan, Luyang Wei, Gangyi Chen, Lin Peng
South China Institute of Software Engineering, Guangzhou University SISE, Guangzhou, China

*Corresponding author e-mail: zh@sise.com.cn

Abstract. Intelligent question-answering system (IQAS) can help users quickly retrieve the answers to the required questions which is currently a popular research direction of NLP. This paper combines the research of relevant scholars and the analysis of their advantages and disadvantages, and an intelligent question-and-answer system is designed, which can automatically generate a question-and-answer knowledge base based on a given document. It can retrieve the knowledge base and recommend answers automatically, quickly and accurately according to the keywords provided by users. The system realized text semantic understanding and analysis, and its MRR value reached a satisfactory 0.7381. At the same time, the paper extends an automatic answer text generation technology based on Seq2Seq model, which can be a useful supplement to the traditional question-and-answer recommendation generation strategy.

1. Introduction
Question-answering (QA) [1-2] is one of the popular information retrieval solutions for current research. In recent years, with the rapid development of the era of big data, more and more information filled with people's lives, and information has become one of the most important resources in the contemporary society.

For most current question answering system [3-5], when users retrieve information, they need to condense their query into several concise keywords then submit them to the question answering system for field search., but with the crazy growth of Internet data, the disadvantage of this kind of question answering system appeared gradually, the vast majority of the user's information demand has not met, people can't fast and exact query from huge data resources to the information you need. However, with the crazy growth of Internet data, the shortcomings of this question-answering system gradually appear. The information needs of most users have not been satisfied, and people cannot quickly and accurately query the required information from the huge data resources. For the current Internet search engine [6, 7], is also the user-supplied keywords later returned to the user information, such as: "NLP is a sub domain of artificial intelligence, search engines will return user many contain keywords" NLP ", "artificial intelligence", "sub areas" of the text, let users to find out in the relevant text, for the average user, they are often difficult to use a small amount of key words to accurately describe the query intent. In addition, this retrieval method returns the result is not a concise and accurate answer, but a list of web page fragments, segments of these web pages usually contains a lot of noise data, users still need to read a large amount of noisy text data to find the answers they need, which makes it a research problem for
people to find the required information accurately and quickly in the era of big data explosion. Therefore, it has become one of the most popular research questions [8-10] to build an intelligent question answering system to enable users to obtain the information they are looking for in the shortest time.

Since the TREC conference held in 1999, it has triggered the research trend of text retrieval [11]. However, due to the insufficient computing power of devices at that time, deep learning and self-access language processing fell into a low period. With the development of the times, the field of computer has developed rapidly. Natural language processing has entered the field of vision with artificial intelligence, and automatic question, as a result, automated question-and-answer computing has grown rapidly. MULDER question-answering system [12] retrieves the network resources based on the keywords provided by users, downloads the retrieved resources to the local place, and processes these documents, thus extracting answers from them. Baidu company research and development of interactive automatic question answering system (baidu know), is from the user's question and answer exchange as a huge online q&a knowledge base, when new problems have been put forward, a searchable database of the system directly, return the highest similarity as the matching problem, and return the answer directly, and other higher similarity of sequential q&a as recommended, in the meantime, the Answer Bus system [13] will return users with a document or long text containing answers. However, such a system has some disadvantages, such as lack of semantic understanding, return a large amount of noisy text, or slow retrieval efficiency.

The IQAS studied in this paper is a set of question recommendation strategies for industry application, which is applicable to technology industries such as large-scale knowledge processing, natural language understanding, knowledge management and automatic question answering system. Through the use of artificial intelligence technology based on the data set (e.g., product manuals, case document, user guide, etc.) to extract knowledge and automatically generate the q&a knowledge base, and effective use of the knowledge base to achieve more accurate and fast problem recommended strategy, implement a current with the industry, but has more efficiency, accuracy, intelligent, flexible, innovative intelligent question answering system. At the same time, the innovation of this system is as follows: at present, most of the industry adopts manual construction of knowledge base. This work provides a frame-based rule feature extraction algorithm, which can automatically extract QA pairs from technical documents and build knowledge base automatically. At present, most of the problem recommendation in the industry adopts the field index method, with single matching, which is not flexible and intelligent enough. This system can realize natural language understanding, and can handle questions in different ways to achieve more accurate problem recommendation. In addition, this study extends the Seq2Seq model to provide some reference for the real q&a.

2. intelligent question and answer system modeling

2.1. Question and answer knowledge base modelling and implementation

Knowledge base is one of the key competitiveness of intelligent customer service/IQAS. At present, most of the knowledge base (q&a) construction of intelligent customer service in the industry is built manually. A complete knowledge base construction requires a lot of manpower. There is an urgent need for an automation solution that can automatically build a knowledge base based on a given document, etc. In this section, a solution strategy is designed for this problem. Meanwhile, Tencent cloud, Huawei cloud and Aliyun technology documents [14-16] are used for question and answer extraction test.

2.2. Quick extraction algorithm of block-type q&a

The algorithm will locate the floor through the similarity analysis of the floors, find some effective information by filtering noise vocabulary, compare its position in the Dom tree, and locate the effective information accurately., through the use of open source library Beautiful soup [17] combined with the regular expression used to screen page block, according to the properties of the HTML DOM tree and HTML tags, considering the particularity of part of the page, and then partitioned based on removing
noise, to regular to design characteristics, through the combination or the combination of titles processing, extract the QA for the problem, The process is as follows:

**Figure 1. Flow chart of quick extraction algorithm for block q&a**

According to figure 1, the input data for the page source or technical documentation, first on the analysis of the input data input data to make use of open source framework Beautifulsoup HTML parsing the DOM tree completion or parsing, through unsupervised word frequency statistics design noise thesaurus to input data denoising, mining rules using regular design features to extract of problem, and then determine the input data structure, to locate the answer to the question, the use of noise filtering method and label characteristic method to extract the answer, through the training of test documents the rules can be expanded, thus improve the versatility of the algorithm.

2.3. Experiment and analysis

Experimental data source for China's three big cloud computing the manufacturer's technical documentation, test results (table 1), the data of the extraction rate of more than 99%, and can extract the file basic all is meaningless without QA for page, shows the superiority of the algorithm, in order to ensure no repeat, in the knowledge base to set up a unique index, according to this framework, and maintenance of a thesaurus for noise and expand and extend the algorithm of key database are rules and rules as long as enough perfect in theory, the universality of the algorithm will be unlimited, and because is matching rules, The time and space complexity of the algorithm are O (n), and the performance of the algorithm is excellent.
Table 1. QA extraction test results for given technical documents

| Test Set    | Number of actual pages | Extract the number of pages | Quantity of Q&A extracted |
|-------------|------------------------|----------------------------|--------------------------|
| Tencent Cloud | 7190                   | 7189                       | 13687                    |
| Aliyun      | 5817                   | 5817                       | 10972                    |
| Huawei Yun  | 3501                   | 3480                       | 6745                     |

3. Problem recommendation strategy modeling

3.1. Question and answer knowledge base modeling and implementation

This paper adopts a word vector calculation method based on cosine similarity, which calculates the spatial complexity of text similarity and the time complexity is far lower than the current word vector construction method in the industry. The idea is to extract all the key words of the question Q (i) and answer (j) respectively. Remove duplicate keywords and make a keyword set KeyWord {ALL (i, j)}. For the keywords of Q (i) and ANSWER (j), the number of occurrences in the statistics is 0, then the set KeyWord {Q (i)} and KeyWord {ANSWER (j)} is obtained.

The vector cosine distance formula between Q (i) and ANSWER (j) is shown in formula 1:

\[
\cos(\text{KeyWord}\{Q(i)\}, \text{KeyWord}\{\text{ANSWER}(j)\}) = \frac{\sum_{i=1}^{n} (x_i, y_i)}{\sqrt{\sum_{i=1}^{n} (x_i)^2} \times \sqrt{\sum_{i=1}^{n} (y_i)^2}}
\]

Description of conformity:

Xi is single numerical element of KeyWord {Q (i)} in the set of problem Q (i).
Yj is the single numerical element of KeyWord {Q (i)}, the set of single elements in answer set WA (i) corresponding to question Q (i).
N is the number of elements in the keyword set KeyWord {ALL (i, j)}.

3.2. Similarity of short text based on Skip-gram word vector

Skip gram model [18, 19] is actually very similar to the idea of auto-encoder in the whole modeling process, that is, a neural network is constructed based on training data first, and its modeling steps are as follows:

1) Sampling

During the modeling process, many common words such as "the" and "a" will appear in the training text (also stop words), which will bring a lot of noise to the training of the model.

Formula 2 is used to calculate the probability of each word being deleted:

\[
p(w_i) = 1 - \frac{t}{f(w_i)}
\]

Where represents the occurrence frequency of word w_i. T is a threshold, generally between 1e-3 and 1e-5. By calculating the probability of each word in the sample being deleted, and sampling based on the probability, the list of sampled words is obtained.

2) Construction batch

Skip -gram is based on an input word to predict the context, so one input word will correspond to multiple contexts. The steps are as follows:
3) Model construction

After the data preprocessing is completed, the model needs to be built. In order to speed up the training and improve the quality of the word vector, the model uses negative sampling to update the weight.

4) Input layer to embedded layer

The weight matrix from the input layer to the hidden layer is used as the embedding layer to be given its dimensions. The embedding size is generally set to between 50 and 300.

5) Embed the layer into the output layer

Multiple contextual words for each input word are actually a Shared weight matrix, with each (input word, input word) training sample used as input. Epochs, batch size and window size were set by adjusting parameters. In order to speed up the training and improve the quality of the word vector, negative sampling was used to update the weight.

After the training, this paper uses the vector cosine distance to calculate the word vector formed by the whole sentence, and then obtains the similarity of the two texts.

3.3. Problem collaborative recommendation strategy

Space vector cosine algorithm calculation speed, but can't be semantic understanding, and Skip-"gram algorithm can carry out semantic understanding, but the computation speed slow, therefore, this article uses the former work out the problem and the problem of knowledge database similarity, similarity of the highest sort after 100 questions, and then use the latter to question similarity calculation, sorting out the highest TOP-N problems are recommended.

3.4. Experiment and analysis

● Experimental data and environment

In order to test the performance of the word vector model proposed above, the experimental data were extracted from 13,687 QA pairs (Data Test) extracted by the rapid extraction algorithm through the block-type q&a proposed in the first chapter according to the tencent cloud technology document, and have been de-duplicated and saved in the mogodb database. The experimental machine is configured with a CentOS 7.3 64-bit version of the operating system, with 1 core CPU and 2 GB of elastic cloud server memory.

● Evaluation indexes

MRR (Mean reciprocal rank) [20] is an internationally accepted mechanism to evaluate the search algorithm, and its calculation formula is as follows:

\[
MRR = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{r_i}
\]

(3)

‘N’ is the number of all answers in the test set, and ‘ri’ is the position of the first correct answer of the ‘i’ question.

● Analysis of experimental results
Table 2. Performance test results of three problem recommendation policies under specified devices

| Question recommendation strategy | spend time(s) | MRR   |
|----------------------------------|---------------|-------|
| SCV                              | 0.62          | 0.5734|
| Skip-gram                        | 89.6          | 0.7574|
| SCV-SG                           | 0.73          | 0.7387|

As can be seen, the computing speed of SCV is relatively fast, but the MRR value is relatively low, while the Skip gram computing speed is relatively slow. However, because of its semantic understanding, the MRR value is relatively high, while the SCV-SG collaborative recommendation algorithm integrates the advantages of the two algorithms, and solves the problem of computing time and accuracy of q&a recommendation well.

4. recommend policy extensions

4.1. Automatic generation of QA pairs based on Seq2seq

Seq2seq model, namely Sequence to Sequence [21, 22]. It can be that the input and output sequences are unequal. The classical RNN model fixes the size of the input and output sequences, while the seq2seq model breaks through this limitation. It realizes the transformation from one sequence to another. For example, Google has implemented the translation function with seq2seq model and attention model. Therefore, this paper USES this method to generate QA pairs automatically, and then realizes the model of generative automatic question answering system.

The main idea of Seq2Seq is to first encode the input sequence into a vector, and then analyze the output sequence by extracting the information in the vector. Seq2Seq structure is composed of encoder and decoder. Encoder is used to encode the input sequence, while decoder is used to parse the output sequence. The network structure of our generated automatic question answering system model is composed of an infrastructure based on Seq2Seq, and its structure is shown in the figure.

![Figure 2. Network model structure](image)

On the network structure, the basic seq2seq structure is adopted, in which GRU [11] is adopted as the coder and decoder. On the network parameters, the dimension size of the word vector is 50, the hidden layer neurons of the GRU is 40, the sample number of batch training is 6000 data, and the learning rate is 0.001. In the training process, the reverse propagation algorithm and Adam optimization algorithm are used. The initialization of all parameters is normally distributed.

4.2. Experiment and analysis

In question-answering task, evaluation standard is generally generated verification deviation between the answer and the standard answer, so we take the front for QA to extract Huawei cloud data set as the training set, 90% 10% as the test set, use 5 layer cross validation method, vector mean square error as the evaluation standard, the automatic question answering system to generate the answer to the sequence...
and the sequence of the standard answer exactly match ratio, the test results in the following table, the model for this essay has certain benefits, along with the length of the text, the MSE sharply reduce, so for this model need more training data and deep research, However, the advantages are also obvious. The performance of the algorithm has certain application benefits. We expect someone to expand the model and achieve the effect similar to Google translation.

Table 3. Seqseq’s test results

| Test Set(question)           | MSE    |
|-----------------------------|--------|
| All characters              | 63.47% |
| No more than 30 characters  | 72.37% |
| No more than 15 characters  | 76.16% |

5. summarizes

In this paper, the intelligent question answering system q&a knowledge base construction, the research achievements of word vector modeling and recommend strategies to systematically expounded in detail. In this paper, the design of intelligent question answering system based on a given document, extract knowledge and automatically generate the q&a knowledge base, avoid the waste of labor, artificial building knowledge base based on knowledge base, use the cosine vector space model and the Skip-gram model two kinds of collaborative computing text similarity word vector model problem recommended, retrieval results not only realized the semantic understanding, and fast retrieval rate, improvements to the existing question answering system, search engine has strong practical value, At the same time, an extended technology based on Seq2Seq model to automatically generate the answer text is proposed. Although this method cannot achieve the expected effect due to the lack of research funds and energy, it has certain reference value for the construction of the automatic question answering system.

References

[1] Bolelli L, Ertekin S, Zhou D, et al. K-svmmeans: A hybrid clustering algorithm for multi-type interrelated datasets [C]. 2007: 198 - 204.
[2] Zheng S F, Liu T, Qin B, et al. Overview of question answering [J]. Journal of Chinese Information Processing, 2002, 16 (6): 46 - 52.
[3] Li S, Zhang J, Huang X, et al. Semantic computation in a Chinese Question-Answering system [J]. Journal of Computer Science and Technology, 2002, 17 (6): 933 - 939.
[4] Sagara T, Hagiwara M. Natural language neural network and its application to question-answering system [J]. Neurocomputing, 2014, 142 (1): 201 - 208.
[5] Jovita, Linda, Hartawan A, et al. Using Vector Space Model in Question Answering System [J]. Procedia Computer Science, 2015, 59:305-311.
[6] Verheggen K, Raeder H, Berven F S, et al. Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows [J]. Mass Spectrometry Reviews, 2017.
[7] Sánchez, David, Martinez-Sanahuja, Laura, Batet M. Survey and evaluation of Web search engine hit counts as research tools in computational linguistics [J]. Information Systems, 2017:S0306437917303290.
[8] Kang L, Jun Z, Shizhu H, et al. Question Answering over Knowledge Bases [J]. IEEE INTELLIGENT SYSTEMS, 2015, 30 (5): 26 - 35.
[9] Li J, Sun Y. [Lecture Notes in Computer Science] Web-Age Information Management Volume 9098 || A Question Answering System Built on Domain Knowledge Base [J]. Lecture Notes in Computer Science, 2015, 10.1007/978-3-319-21042-1(Chapter 9): 111 - 122.
[10] Lin Y, Shen H. SmartQ: A Question and Answer System for Supplying High-Quality and Trustworthy Answers [J]. IEEE Transactions on Big Data, 2017: 1 - 1.
[11] Voorhees E M, Hardman D K. The {Ninth Text Retrieval Conference} (TREC-9) [J]. Information
[12] C. Kwok, O. Etzioni, and D. S. Weld. Scaling Question Answering to the Web [A]. In: Proceedings of the 10th World Wide Web Conference (WWW2001) [C], Hong Kong, 2001, 150 - 161.

[13] Brill E, Dumais S, Banko M. An Analysis of the AskMSR Question-Answering System [C]// Acl-02 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2002.

[14] https://cloud.tencent.com/?fromSource=gwzcw.2212127.2212127.2212127&utm_medium=cpd&utm_id=gwzcw.2212127.2212127.2212127.

[15] https://www.aliyun.com/utm_content=se_1000301881.

[16] https://activity.huaweicloud.com/2019Feb-promotion/index.html?utm_source=Baidu&utm_medium=brand&utm_campaign=HWC_PC&utm_content=Title&utm_term=%E5%BC%80%E5%B9%B4%E9%87%87%E8%B4%AD%E5%AD%A3

[17] Nair V G. Getting Started with Beautiful Soup [J]. 2014.

[18] Henry S, Cuffy C, Mcinnes B T. Vector representations of multi-word terms for semantic relatedness [J]. Journal of Biomedical Informatics, 2018, 77: 111.

[19] Balikas G, Amini M R. An empirical study on large scale text classification with skip-gram embeddings [J]. 2016.

[20] https://baike.baidu.com/item/MRR

[21] Chung E, Park J G. Sentence-Chain Based Seq2seq Model for Corpus Expansion [J]. Etri Journal, 2017, 39 (4): 455 - 466.

[22] Strobelt H, Gehrmann S, Behrisch M, et al. SEQ2SEQ-VIS: A Visual Debugging Tool for Sequence-to-Sequence Models [J]. IEEE Transactions on Visualization and Computer Graphics, 2018: 1 - 1.