Automatic Question Answering System for Semantic Similarity Calculation

MinChuan Huang, Ke Chen, XingTong Zhu, and GuoQuan Wang

Abstract The automatic question answering system based on semantic similarity calculation includes three modules: word segmentation module, question understanding module and FAQ database module. Jieba, an open-source tool, is used in the word segmentation module. The problem understanding module can be further divided into problem classification, keyword extraction, and keyword expansion. The hierarchical classification method based on self-learning rules is used for problem classification. The common question database module distinguishes sentence similarity calculation and question matching. Sentence similarity calculation is based on the How Net semantic dictionary. The core algorithm is the rule algorithm design based on the corpus. The system relies heavily on each module, so it is difficult to establish a more perfect test scheme. Therefore, we only test the sentence similarity calculation which ultimately determines the accuracy of the problem matching, and finally realize the function of each module, and test and evaluate each module. The test results can be summarized that the sentence segmentation is relatively short, the part of speech contains less, and the similarity judgment is relatively concentrated, which is caused by the absence of specified parts of speech in both sentences. According to the part of speech coverage specified by the system, the more comprehensive the coverage, the more accurate the similarity calculation.

Keywords Machine learning · Chinese word segmentation · Question classification · Semantic similarity · Question base

1 Introduction

This is a newly developed Chinese question answering system [1–3], covering computer linguistics, machine learning [4–6] and natural language processing technology, code implementation, algorithm principle of the system, artificial intelligence and information retrieval [7]. System and design of the thinking scheme,

M. Huang · K. Chen · X. Zhu (✉) · G. Wang
Guangdong University of Petrochemical Technology, Guangdong 525000, China
e-mail: 305299282@qq.com

E. C. K. Cheng et al. (eds.), *Artificial Intelligence in Education: Emerging Technologies, Models and Applications*, Lecture Notes on Data Engineering and Communications Technologies 104, https://doi.org/10.1007/978-981-16-7527-0_7

natural language questions, semantic similarity analysis, question answering system can automatically answer. In order to improve the efficiency of information retrieval when answering questions, semantic similarity calculation is needed to accurately match a large number of existing questions in the question database, find out the appropriate answers, and give feedback to users.

The way to retrieve information on the Internet is to type keywords through the search engine. The traditional search engine mode is that users return a large number of hyperlink indexes of relevant pages by typing keywords. Users filter relevant pages through the introduction of the page and the matching of keywords, and then enter the relevant pages to browse and further filter the information they want [8].

Advanced automatic question answering system mode, the user can directly input questions in natural language, and the system can return the user's desired answers in natural language by understanding the questions. As a more advanced information retrieval technology, automatic question answering system has been put forward and received widespread attention. Users can ask questions in natural language and get accurate answers quickly through automatic question answering system [9].

The main research institutions of Chinese retrieval information are Institute of computer science, Chinese Academy of Sciences, Fudan University, Harbin Institute of technology, Beijing Language and Culture University, etc. The Chinese Academy of Sciences has developed National Knowledge Infrastructure (NKI) question and answer system. The knowledge base includes more than ten knowledge bases, such as geographic knowledge, weather forecast, character knowledge, etc. In terms of Chinese word segmentation technology, there are nlpir of Chinese Academy of Sciences, LTP of Harbin Institute of technology, and thulac of Tsinghua University.

2 System Analysis and Theoretical Conception

The theoretical basis of Chinese automatic question answering system is different from English grammatical structure [10]. Chinese sentence structure is highly complex. It is difficult in natural language processing technology. How to make word segmentation fast and accurate is the primary problem of Chinese automatic question answering system. There is also the problem classification, which needs a deeper analysis of sentences. Sentence similarity calculation and question matching are based on the result of sentence similarity calculation. If the similarity is higher than a certain value, the matching is considered successful. There are also two problems, ambiguity resolution and unknown word recognition.

2.1 Chinese Word Segmentation Tool

Among the existing Chinese word segmentation tools [11], the Natural Language processing & Information Retrieval Sharing Platform (NLPIRSP) [3, 12] of the

Institute of computing technology, Chinese Academy of Sciences can fully meet the processing needs of users, including network crawling, text extraction, Chinese word segmentation, word tagging and so on. It is the most comprehensive text processing tool in the current word segmentation tools. At present, among the Chinese word segmentation tools, the most comprehensive text processing tool is the Natural Language Processing & Information Retrieval Sharing Platform (NLPIRSP) of Institute of computing technology, Chinese Academy of Sciences. NLPIRSP can meet the needs of users for big data text processing from all aspects, including the complete technical chain of big data: network capture, text extraction, text retrieval, text retrieval, etc. Chinese and English word segmentation, part of speech tagging, entity extraction, word frequency statistics, keyword extraction, semantic information extraction, text classification, emotion analysis, semantic depth expansion, complex and simple coding conversion, automatic phonetic notation, text clustering, etc. NLPIR provides rich open API, which can be integrated into all kinds of complex operating systems, and can also be called by all kinds of mainstream development languages.

2.2 Chinese Word Segmentation Algorithm

The segmentation algorithm of Chinese dictionary is based on the matching and cutting of Chinese sentences and existing dictionaries. According to the length of matching words, maximum matching and minimum matching are distinguished. Forward matching and reverse matching. Chinese word segmentation faces two major problems, ambiguity resolution and unknown word recognition. The causes of ambiguity can be divided into three types: intersection ambiguity, combination ambiguity and true ambiguity. The identification of unknown words is not included in the dictionary, which can be divided into proper nouns and non-proper nouns. With the change of social life, there will be more and more categories of unlisted words.

The main purpose of question classification is to classify according to the types of questions and improve the efficiency and accuracy of question answering system. Question classification can effectively simplify the classification of candidate answers. Bayes theorem formula, P (H|x) = (P (x|h) P (H))/P (x), where p (H|x) belongs to a posteriori probability, which refers to the probability of h under condition X. P (H), P (x) and P (x|h) belong to a priori probability, and the former two refer to the probability of X and H events. The latter refers to the probability of X under H condition. Naive Bayes algorithm is used to adjust Bayes classification method, and its simple formula is shown in (1) below [13, 14].

$$\mathrm{P}(\mathrm{X}|C_i) = \prod_{n}^{k=1} P(x_k|C_i) \tag{1}$$

3 System Module Design

Chinese automatic question answering system is divided into user view interface and question processing background. The user view interface is used to give the user feedback of typing questions and getting answers [15, 16].

The background of question processing includes three parts: word segmentation module, question understanding module and quick Q & A library module. The structure of Chinese automatic question answering system is shown in Fig. 1.

3.1 Chinese Word Segmentation System

Chinese word segmentation adopts Jieba word segmentation [17, 18] of Python open source tool. Jieba word segmentation tool provides user dictionary interface, which is conducive to the expansion of subsequent professional vocabulary, and realizes more accurate word segmentation of professional vocabulary. Jieba word segmentation is based on trie tree structure to achieve efficient word graph scanning and generate all possible Directed Acidic Graph (DAG). The word segmentation tool has a dictionary, which contains more than 20,000 words. The word structure includes the word itself, word frequency and part of speech, and word frequency statistics. Before word segmentation, the dictionary is loaded in advance and loaded into a Trie tree to improve the search speed.

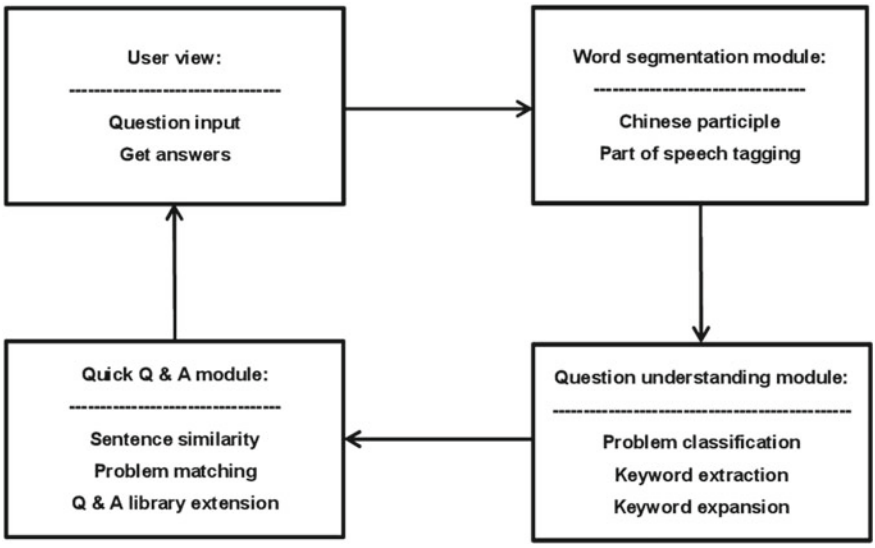

Fig. 1 The structure of Chinese automatic question answering system

Table 1 Styles available in the word templet

Code	Noun	Annotation
a	Adjective	The first letter
n	Noun	The first letter
v	Verb	The first letter
d	Adverb	The second letter
m	Numeral	The third letter
nr	Name	The first letter

For the recognition of the non-logins, the Jieba participle uses Viterbi algorithm based on the Hidden Markov Model (HMM) model to realize the prediction segmentation. HMM is a mathematical statistical model, which is used to describe the HMM process with unknown parameters. Through the word frequency record in the dictionary, the Jieba segmentation uses dynamic programming to find the maximum probability path, finds the maximum partition combination, and solves the problem of ambiguity resolution. The words are divided into the best, the frequency of words is determined, the maximum path is determined according to the dynamic programming method, and the maximum probability of segmentation combination is obtained.

Part of speech tagging, Jieba word segmentation tool supports part of speech tagging at the same time. Part of speech tagging adopts the part of speech tagging set of Peking University Institute of computing. Part of speech tagging is helpful to understand the structure of questions, and it is the main parameter for the subsequent problem classification and sentence similarity processing. See Table 1 for the code, noun and annotation of part of speech tagging.

3.2 Problem Classification and Keyword Generation

The hierarchical classification method based on machine learning rules is used to realize Chinese problem classification [19]. The training result of QCR and QHCR rule matching is incomplete, so it is not used. Bayesian classifier is selected to improve the classification accuracy, keyword extraction and keyword expansion. Figure 2 flow chart of problem classification.

Keyword extraction, input question segmentation, need to extract the key words in the question. We need to filter some words with low information content, extract the words with high information content, and finally use the remaining keywords for subsequent similarity calculation.

In order to identify synonyms and many uncommon professional words, the system needs to expand keywords.

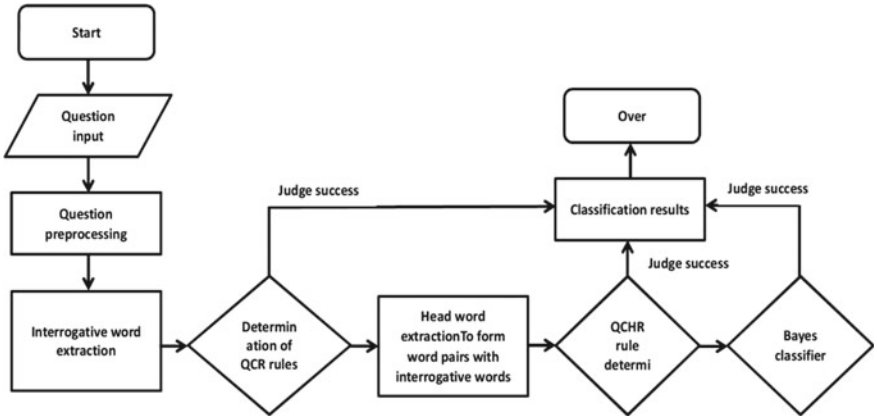

Fig. 2 Flow chart of problem classification

3.3 Sentence Similarity Calculation

It is divided into sememe similarity calculation, word semantic similarity calculation and sentence similarity calculation. The word similarity calculation based on HowNet is adopted, and the relevant sememes are extracted as the basis of similarity matching through the corresponding knowledge structure of each word in HowNet dictionary.

Sememe Similarity Algorithm

For the synonymous primitive tree, according to the characteristics of the tree structure, the distance between the two sememe nodes is calculated, and then the similarity value of the sememe is calculated through the formula. As shown in Fig. 3, Y1 and Y2 traverse upward to find the common cross node Y7, and the sum of their steps is the distance between the two primitive nodes.

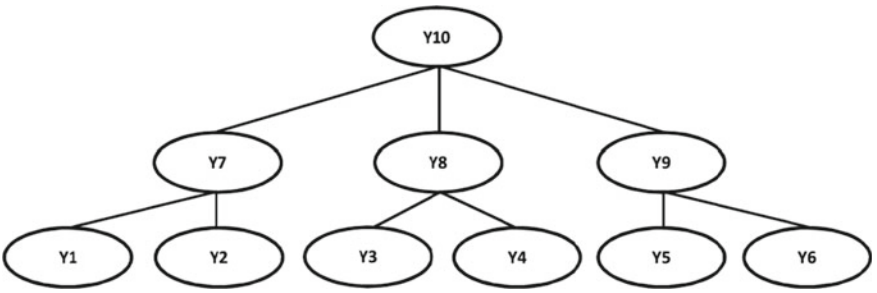

Fig. 3 The tree structures

Word Similarity

In China National Knowledge Infrastructure (CNKI) dictionary, every dictionary will have a sememe description, including one or more sememes. The similarity of each sememe is calculated by rules, and the word similarity is calculated by weighted conversion.

Sentence Similarity

The calculation of sentence similarity is based on the calculation of word similarity. Through the word segmentation module, the cut set of words is obtained. According to the importance of words in the sentence, the weighted calculation is carried out to get the sentence similarity value.

Figure Descriptions

Every figure should have a figure description unless it is purely decorative. These descriptions convey what's in the image to someone who cannot see it. They are also used by search engine crawlers for indexing images, and when images cannot be loaded.

Problem Matching

The method of question classification and sentence similarity calculation matches the question set in the fast question answering database, and the answers are responded to the users. Questions in question answering database will be classified and keywords extracted in advance, and the results will be loaded into memory to improve the efficiency of question matching.

4 Core Algorithm

The system uses Python language which is open source on the Internet to develop Jieba word segmentation component of Chinese word segmentation module. The function of Jieba word segmentation module includes word segmentation, part of speech tagging and keyword extraction. The keyword extraction of Jieba word segmentation is based on the TF-IDF algorithm and text rank algorithm, which calculates the weight of words in a large number of texts. The corpus of question answering corpus of this system mainly exists in the form of short questions, and there is no data base of a large number of texts. Therefore, the key word extraction function of Jieba word segmentation is not used [20, 21].

4.1 Word Segmentation Module

There are styles for block quotations, which should be used for quotes that are separated from in-line text. Below is an example.

Jieba word segmentation function and part of speech tagging function to achieve the design of word segmentation module. According to the question and answer corpus of this system, all the existing questions are directly segmented with part of speech, and the segmentation results are analyzed.

Jieba word segmentation is based on its own vocabulary. The parts of speech of punctuation and unmarked words are marked with X; There are no English words in the dictionary, and the English part of speech is marked as end, which is compatible with non-Chinese words and symbols in the text. Jieba word segmentation component is enough to meet the basic needs of daily sentence segmentation. The system is a question answering system for the information field. All questions contain words in the information field, which is different from the results of daily word segmentation.

4.2 Problem Classification

To classify problems, we need to determine the type system of problems. There is no unified classification system for Chinese problem classification. This system establishes its own problem classification system based on the characteristics of information field. According to the UIUC classification system standard and the characteristics of Chinese problems, Harbin Institute of technology adopts the hierarchical classification method to formulate the Chinese problem classification system.

This classification system is directly used as the basis of the system problem classification system to ensure that the system can better expand the problem classification system. To classify problems, we must first establish the type system of problems.

The question and answer set in information field focuses on DES (description class). Des (description class) includes reason, relation, effect, form, behavior, rules, comment, purpose, trait, type and advantage. See Table 2 problem classification system.

After establishing the classification system, we need to establish the problem classification rules. The hierarchical classification method based on self-learning rules can reduce the workload of rule establishment and the subjectivity of human operation in classification. To establish self-learning rules, we need to extract interrogative words and head words first.

Interrogative words include general interrogative words and special interrogative words. Special interrogative words are used to classify problems, which can be realized by one interrogative word. General interrogative words are used to classify problems when there are not enough interrogative words. When the question contains general interrogative words, it needs the auxiliary judgment of head words.

Table 2 Problem classification system

Classification	Details
DES (description class)	Adventage, Behavior, Comment, Charact, Effect, Form, Purpose, Reason, Relation, Rules, Type, Other
HUM (people)	Description, Organization, Person, Other
LOC (location)	Address, Country, Other
NUM (digital)	Area, Code, Count, Frequency, Percent, Range, Speed, Temperature, Other
OBJ (entity class)	Academic, Event, Food, Instrument, Language, Substance, Other
Time	Era, Time, Other
Unknown	

The head word reflects the essence of the problem and does not have ambiguity. The combination of general interrogative words and head words can classify questions.

To determine the role of interrogative words and head words in problem classification, it is necessary to design self-learning rules for problem classification. Self-learning rules can be divided into two parts, one is interrogative word type self-learning rule (QCR rule), the other is interrogative word head type self-learning rule (QHCR rule).

Self-learning Rules of Interrogative Words

In order to facilitate the expansion and adjustment of the interrogative words, the interrogative words are extracted from the self-learning rules, and the way of establishing the interrogative words list is to extract the interrogative words from the sentences. The purpose of learning rules is to determine the association rules of question words and question types, extract question types, and realize the manual division and annotation of question sentences in training corpus.

Self-learning Rules of Interrogative Head Words

The self-learning rule and the self-learning rule of interrogative word type add a parameter to match the head word. It is necessary to establish this rule to match the general interrogative words that cannot be matched. The extraction of head words is more difficult than the extraction of interrogative words. We can build an interrogative word list to match interrogative words, but we can't build a head word list because the number of head words is more than that of interrogative words, and there are many possibilities of matching with general interrogative words. The existing algorithms for the extraction of head words include syntactic dependency structure analysis. The implementation of syntactic dependency structure is complex and needs a large number of training corpus, so it is not used.

Improved Progressive Bayesian Classification Method

Self-learning rules can improve the accuracy of problem classification to a limited extent, but the rules extracted from the head words cannot cover all types of problems, which is caused by the complexity and diversity of Chinese question structure. The solution is to avoid the problems that cannot be covered by self-learning rules cannot be classified, import the third level classification rules, and improve the Bayesian classification method.

The first step of the system is Chinese word segmentation. The cut words are filtered by stop word list, and the remaining words are calculated as eigenvalues. The eigenvalues of Bayesian model should be independent as far as possible. Assuming that the problems in the problem set conform to the bag of words model, the remaining words after word segmentation are independent of each other, the position and order are not related, and some structural and semantic information will be lost. However, it can improve the algorithm implementation of Bayesian model. Set 0.5 as the zeroing factor, n represents the total number of problem types, and the value of n is 34. Combined with TF-IDF for weight processing, it needs to reduce the weight of words manually. X is the number of questions contained in the related question type, and Y is the number of times that the word appears in the related type.

$$P_2(qc, Q_1) = P_1\left(qc_i, Q_i\right) \times \log\left(\frac{N + 0.1}{N + 0.1}\right) \tag{2}$$

4.3 Keyword Extraction

Questions are segmented and keywords are extracted. The speed and accuracy of question matching can be improved by eliminating the low information words that hinder sentence similarity matching. The stop words database developed by Harbin Institute of technology is directly used as the basis for the establishment of stop words list. The low information words are filtered and the stop words list is expanded and adjusted.

The key points of the system include noun as part of speech, extended vocabulary of noun, adjective, verb and restrictive adverb. This system can't carry out this kind of division, and give different weights according to the part of speech for sentence similarity matching. Math statements should have the "Statement" style applied.

4.4 Word Similarity Calculation

After keyword extraction, sentence similarity matching becomes the similarity calculation of keyword set. Every computing unit is in the form of words, the first solution is the word similarity calculation.

This paper analyzes the structure of sememe description and divides it into four parts: the first independent sememe description, other independent sememe description, relational sememe description and symbolic sememe description. The word similarity calculation is divided into four aspects.

The first independent sememe description is calculated and its similarity is recorded as Sim1 (Y1, Y2). The similarity of other independent sememe descriptors was recorded as sim2 (Y1, Y2). All the independent sememes (except the first one) of two expressions are arbitrarily paired, and the sememe similarity of all possible pairings is calculated. The one with the highest similarity is selected, and they are grouped into one group. This process is repeated until all the independent sememes are grouped. The similarity is recorded as sim3 (Y1, Y2). And then calculate their similarity. The similarity of sememe description is calculated and recorded as sim4 (Y1, Y2). The calculation method of symbolic sememe description is consistent with that of relational sememe description. Through the above four parts of similarity calculation, weighted calculation can get the final similarity value of the two words. So far, the word similarity calculation is realized, which lays the foundation for the subsequent sentence similarity calculation.

$$\mathrm{sim}(w1, w2) = \sum_{i=1}^{4} \beta_i \mathrm{sim}_i(Y1, Y2) \tag{3}$$

4.5 Sentence Similarity Calculation

The final result of sentence similarity calculation affects the accuracy of problem recognition and matching. The following rules should be established to improve the efficiency and accuracy of sentence similarity calculation, and finally calculate the similarity value of sentences by combining the two.

The establishment of professional vocabulary list can improve the accuracy of the recognition of professional vocabulary by Jieba participators, and mark the word quality of professional words. The segmentation of Jieba has the ability to mark the word character. Only two words with the same word quality are calculated to improve the recognition efficiency of the system.

After the word and keyword extraction, we can get the key words set of two sentences. The first rule only calculates the similarity of words consistent. Match the words that meet the conditions, construct the matching matrix, calculate all similarity

values, and take the maximum value for each line of the matrix, which is the similarity value of a keyword in the question to be matched.

When one word exists in one question and the other does not exist, the similarity value is defined as a small constant.

To determine the similarity value of key words in a sentence, we should make weighted calculation to obtain the final sentence similarity value. The formula is established: sentence similarity = noun similarity *a1 + verb similarity *a2 + adjective similarity *a3 + adverb similarity *a4 + other words *a5. After repeated statistics, the final value of each parameter is a1 = 0.3, a2 = 0.3, a3 = 0.18, a4 = 0.18, a5 = 0.04.

4.6 System Implementation Results

After the key words are extracted from the quick Q & A database module, we need to combine the word similarity method of HowNet and the weighted formula to calculate the sentence similarity. For the functional test of sentence similarity, the system test parameters are: the weight of noun part of speech is 0.3, the weight of verb part of speech is 0.3, the weight of adjective part of speech is 0.18, the weight of adverb part of speech is 0.18, and the weight of other parts of speech is 0.04, The similarity of the part of speech is 0.70. If one sentence has a specified part of speech and another sentence does not, the similarity of the part of speech is 0.1.

The test results show that the sentence structure is relatively short, the part of speech contains less, and the similarity judgment is relatively concentrated, which is caused by the system does not have the specified part of speech for both sentences. According to the part of speech coverage specified by the system, the more comprehensive the coverage, the more accurate the similarity calculation.

As shown in Table 3, in the last test case, the query content includes “rules”, “abbreviations” and “hieroglyphs”, and the artificial questions only contain “hieroglyphs”. The system should be able to realize question matching, and the actual similarity value is less than 0.75, so the question matching should be realized by the way of question simplification. From the results, there are still some problems in sentence similarity calculation, which need to be further studied.

What is the content of computer security? What’s the difference between e-mail protocol and e-mail protocol? The accuracy of Q & A similarity reached the highest of 0.88. How to construct open intranet? The similarity accuracy is the lowest, only 0.42.

Table 3 Sentence similarity test results

Original question	Artificial judgment	Similarity	Dissimilar sentences	Similarity
What is computer security?	What is the content of computer security?	0.88	What is hardware description language?	0.69
Why check the electrical rules?	What is the purpose of electrical rule detection?	0.80	Why design green chip?	0.61
What are the common email protocols?	What is the email protocol?	0.88	How do secure electronic transactions work?	0.54
How to realize informatization in enterprises?	How should a company realize informatization?	0.74	How to construct open intranet?	0.42
How to see a doctor through telemedicine system?	How can I use telemedicine to see a doctor?	0.82	How to set up campus network?	0.48
How to understand the rules, abbreviations and hieroglyphs of e-mail?	How to understand hieroglyphs?	0.61	How to use search engine?	0.53

5 Conclusions

Based on the core corpus of "the new 100,000 whys of computer and information science", this study constructs a common question database, and establishes a question system of more than 30 categories according to the characteristics of the corpus. Chinese word segmentation and part of speech tagging are realized by using Jieba word segmentation tool. Semantic understanding and similarity calculation are realized through the existing research results of HowNet dictionary. Hierarchical classification method improves the accuracy of rule classification. Python scripting language features, build functional module division, through the way of file construction to reduce the impact of module compatibility.

The development environment of the Chinese automatic question answering system is based on Windows system, combined with PyCharm development environment, using Python 3.6 programming language for development, using QT IDE of QT creator cross platform as interface and user interface (UI) for development. Python language is developed in the form of script. The core algorithm of the system includes word similarity calculation based on HowNet, sentence similarity calculation method based on word similarity, hierarchical classification method based on self-learning rules, etc. it is modular organization based on the characteristics of Python script language.

The success or failure of automatic question answering system depends on the integration of module functions, and its error will affect the decline of system accuracy.

What is the content of computer security? What's the difference between e-mail protocol and e-mail protocol? The highest similarity of Q & A was 0.88. How can I use telemedicine system to see a doctor? 82. What is the purpose of electrical rule testing? The similarity reached 0.80. How should a company realize informatization? The similarity reached 0.74. How to construct open intranet? The accuracy of similarity is the lowest, only 0.42.

This follow-up study can be extended to the wide application of university education administration. For example, when freshmen enter the University, they do not know the procedures and format templates of application documents for the relevant provisions of the student manual. At this time, the automatic question and answer app system of student manual is required to effectively guide the freshmen to handle relevant business. For example, the graduation design (Thesis) automatic question and answer web or app system of fresh graduates can assist the fourth-year students to inquire about the relevant provisions of the thesis format, how to download the relevant templates and templates, and how to handle the graduation and departure procedures.

Acknowledgements We would like to thank the School of Computer Science, Guangdong University of Petrochemical Technology (124400007270402306). Project Number: 2019rc076, 2019rc078. Natural Science Foundation of Guangdong Province: 2018A030307032, 2018A030307038. Key research platform and project of universities in Guangdong Province: 2020zdzx3038.

References

1. Noraset T, Lowphansirikul L, Tuarob S (2021) WabiQA: a wikipedia-based thai question-answering system. Inf Process Manag 58:102431. https://doi.org/10.1016/j.ipm.2020.102431
2. Zihayat M, Etwaroo R (2021) A non-factoid question answering system for prior art search. Expert Syst Appl 177:114910. https://doi.org/10.1016/j.eswa.2021.114910
3. Pesquita C, Faria D, Falcão AO, Lord P, Couto FM (2009) Semantic similarity in biomedical ontologies. PLoS Comput Biol 5:e1000443. https://doi.org/10.1371/journal.pcbi.1000443
4. Chan H-Y, Tsai M-H (2019) Question-answering dialogue system for emergency operations. Int J Disaster Risk Reduct 41:101313. https://doi.org/10.1016/j.ijdrr.2019.101313
5. Zhu S, Cheng X, Su S (2020) Knowledge-based question answering by tree-to-sequence learning. Neurocomputing 372:64–72. https://doi.org/10.1016/j.neucom.2019.09.003
6. Zhao L, Zhang A, Liu Y, Fei H (2020) Encoding multi-granularity structural information for joint Chinese word segmentation and POS tagging. Pattern Recogn Lett 138:163–169. https://doi.org/10.1016/j.patrec.2020.07.017
7. Mohasseb A, Bader-El-Den M, Cocea M (2018) Question categorization and classification using grammar based approach. Inf Process Manag 54:1228–1243. https://doi.org/10.1016/j.ipm.2018.05.001
8. Xiong H, Wang S, Tang M, Wang L, Lin X (2021) Knowledge graph question answering with semantic oriented fusion model. Knowl-Based Syst 221:106954. https://doi.org/10.1016/j.knosys.2021.106954

9. Liu J, Wu F, Wu C, Huang Y, Xie X (2019) Neural Chinese word segmentation with dictionary. Neurocomputing 338:46–54. https://doi.org/10.1016/j.neucom.2019.01.085
10. Yuan Z, Liu Y, Yin Q, Li B, Feng X, Zhang G, Yu S (2020) Unsupervised multi-granular Chinese word segmentation and term discovery via graph partition. J Biomed Inform 110:103542. https://doi.org/10.1016/j.jbi.2020.103542
11. Ferreira JD, Couto FM (20) Multi-domain semantic similarity in biomedical research. BMC Bioinform 20:246. https://doi.org/10.1186/s12859-019-2810-9
12. Natural Language Processing & Information Retrieval Sharing Platform. NLPIRSP Homepage. http://www.nlpir.org/wordpress. Accessed 16 June 2021
13. Ayala J, García-Torres M, Noguera JLV, Gómez-Vela F, Divina F (2021) Technical analysis strategy optimization using a machine learning approach in stock market indices. Knowl-Based Syst 225:107119. https://doi.org/10.1016/j.knosys.2021.107119
14. Yu J, Zhu Z, Wang Y, Zhang W, Hu Y, Tan J (2020) Cross-modal knowledge reasoning for knowledge-based visual question answering. Pattern Recogn 108:107563. https://doi.org/10.1016/j.patcog.2020.107563
15. Zhao S, Wu Y, Tsang Y-K, Sui X, Zhu Z (2021) Morpho-semantic analysis of ambiguous morphemes in Chinese compound word recognition: an fMRI study. Neuropsychologia 157:107862. https://doi.org/10.1016/j.neuropsychologia.2021.107862
16. Li M, Li Y, Chen Y, Xu Y (2021) Batch recommendation of experts to questions in community-based question-answering with a sailfish optimizer. Expert Syst Appl 169:114484. https://doi.org/10.1016/j.eswa.2020.114484
17. Zhenqiu L (2012) Design of automatic question answering system base on CBR. Procedia Eng 29:981–985. https://doi.org/10.1016/j.proeng.2012.01.075
18. Zhang L, Lin C, Zhou D, He Y, Zhang M (2021) A Bayesian end-to-end model with estimatuncertainties for simple question answering over knowledge bases. Comput Speech Lang 66:101167. https://doi.org/10.1016/j.csl.2020.101167
19. Zafar H, Dubey M, Lehmann J, Demidova E (2020) IQA: interactive query construction in semantic question answering systems. J Web Semant 64:100586. https://doi.org/10.1016/j.websem.2020.100586
20. Vanam MK, Amirali Jiwani B, Swathi A, Madhavi V (2021) High performance machine learning and data science based implementation using Weka. Mater Today: Proc S2214785321005617. https://doi.org/10.1016/j.matpr.2021.01.470
21. Xu Y, Zhou Y, Sekula P, Ding L (2021) Machine learning in construction: from shallow to deep learning. In: Developments in the built environment, vol 6, p 100045. https://doi.org/10.1016/j.dibe.2021.100045