Advances in Economics, Business and Management Research, volume 71
4th International Conference on Economics, Management, Law and Education (EMLE 2018)

Research on Chinese Question Answering System in MOOC Environment*

Qiang Qu
Department of Computer Science and Technology
Yanbian University
Yanji, China

Yahui Zhao**
Department of Computer Science and Technology
Yanbian University
Yanji, China
**Corresponding Author

Rongyi Cui
Department of Computer Science and Technology
Yanbian University
Yanji, China

***Abstract*—In view of the lack of immediate question answering system for curriculums in current Chinese MOOC environment, an intelligent curriculum question answering system was designed and applied to MOOC platform in this paper by combining FAQ database and Web. In the stage of question analysis, the user's questions were reasonably standardized using domain dictionaries and N-gram model to effectively detect the errors in terms. Furthermore, web search engine, web crawler and automatic summarization technology were adopted to provide users with recommended answers for the questions which are not logged in the local FAQ database. The test results show that the correction rate of the text proofreading function of the question sentence is 85.3%; the accuracy rate of the answer summary based on web multi-document generation is 78.8%; the accuracy rate of the system is 87.3% and the recall rate is 89.3%.**

Keywords—MOOC platform; question answering system; question text proofreading; multi-document automatic summarization

I. INTRODUCTION

Since 2014, MOOC (Massive Open Online Course) platform has experienced rapid development, which has provided new opportunities for teaching in colleges and universities and online education [1]. The educational mode on the MOOC platform makes our study so convenient that higher education in colleges and universities is available anytime and anywhere [2]. The platform is a massive revolution for teaching methods in colleges and universities [3]. As MOOC platform is still at growth stage, there are still some deficiencies in application. For example, there is no chance for face-to-face communication between teachers and students, and no timely answers or response to students' questions can be provided, which will easily reduce students' enthusiasm in course learning and further cause their misunderstanding or lack of course knowledge. To settle above issues, QA (Question Answering) system that people are following and studying will become a new question answering model in courses [4], which will effectively compensate deficiencies in online education.

An intelligent question answering system has been designed and realized based on the integration of FAQ database and Web technology in this paper to overcome the deficiencies that there is no instant question answering system for courses in current MOOC platform. In proposed QA system, students' questions can be answered by the modules like question analysis, information retrieval and answer extraction. The text proofreading for interrogative sentences is built in question analysis module where answers will be automatically generated from multiple documents retrieved in Web for answer extraction. The system performances have been examined in the practical teaching application and the results are satisfactory.

II. CONSTRUCTION AND PROCESS FLOW OF QUESTION ANSWERING SYSTEM

A. Construction of Question Answering System

The question answering system mainly consists of three modules: question analysis, information retrieval and answer extraction. Question analysis module mainly aims to correctly grasp question intentions in interrogative sentences, and information retrieval module is used to locate document information associated with students' interrogative sentences, and answer extraction module intends to screen out the best answers from document collections of retrieval results [5]. The structure of QA system framework built in this paper is shown as "Fig. 1".

* This research was partially supported by State Language Commission of China under Grant No. YB135-76, Research Topic of Higher Education Teaching Reform of Jilin Province under Grant No. JJGZ 32 [2016].

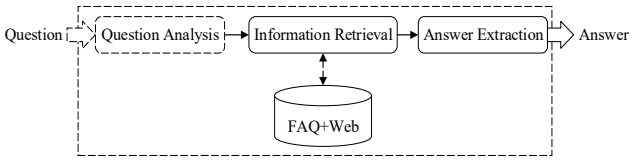

Fig. 1. Structure of QA system framework.

QA system mainly consists of FAQ (Frequently-asked Question) database and Web documents [6]. When students submit their questions, the system will retrieve the answers in FAQ database firstly. If the answers are retrieved in FAQ database, these answers will be sent back to students. Currently, QA system applied for specific areas are mostly realized based on FAQ database with high performance. If the answers are not found in FAQ database, the system will retrieve Web documents containing appropriate answers, and automatically generate answers from multiple documents.

B. *Process Flow of System*

The process flow of QA system constructed in this paper is as follows:

Step 1: Analyze the questions submitted by students. The analysis results include key words in questions, extension for key words, text proofreading for interrogative questions and question filtration that are not associated with course, which will be used in modules afterwards.

Step 2: Pre-process the interrogative questions which have been entered including word segmentation and remove stop words. After acquiring key words from interrogative sentences, the key words will be extended based on synonymy according to Tongyici Cilin and relevant information will be acquired.

Step 3: Carry out text proofreading and filtrating irrelevant interrogative sentences for the final results after pre-process is completed with terminological dictionary.

Step 4: Calculate the similarity value between students' questions and those in FAQ database, and sort the questions in FAQ database based on similarity value.

Step 5: Set up a threshold value for question similarity to determine whether there are similar questions in FAQ database. If similarity value is lower than the threshold value, search for questions through Baidu search engine.

Step 6: Formulate multiple documents by accurate extractions from texts in first five webpages searched.

Step 7: Process the extracted multiple documents, and add the results into database as answers for teachers to review and improve.

Step 8: Correct questions and answers in database by means of teachers' review and modification.

III. Design of Text Proofreading for Interrogative Sentences

Students often make mistakes in professional terms of courses when they submit their questions about courses. The text proofreading for interrogative sentences aims to proofread wrong professional terms in students' interrogative sentences and provide corresponding correction suggestions. The professional terms in the interrogative sentences will be corrected by adopting *N*-gram language model and terminological dictionary in our work. The terminological dictionary consists of two parts. One is computer vocabulary dictionary in the thesaurus of Sogou input; and the other is from manual collection and reorganization, where professional terms of courses will be screened out from different versions of electronic textbooks of *Fundamentals of Computers* manually, which will be processed with word segmentation and part of speech tagging. And then the terms will be examined and corrected by teachers to exclude non-professional terms.

A. *Extraction and Extension of Key Words in Interrogative Sentences*

The questions submitted by students will be pre-processed and key words in the interrogative sentences will be extracted. The process includes word segmentation, part of speech tagging and stop words removing, which steps are illustrated in the following context.

- Stop words removing. The stop words in the students' questions were filtrated by adopting stop words list built by Harbin Institute of Technology.
- Word segmentation and part of speech tagging. The LTP from Harbin Institute of Technology was adopted for word segmentation and part of speech tagging in interrogative sentences.
- Key words extension. Key words extracted are not comprehensive enough for information retrieval module, so key words needs to be extended. The dictionary of Tongyici Cilin (Extended) was adopted in this paper for key words extension in interrogative sentences.

B. *Automatic Text Proofreading for Interrogative Sentences*

Suppose *N*-order Markov Chains is $W_1W_2...W_n$, and the previous *N*-1 elements influence the probability of occurrence of W_i element [7], the probability of occurrence for statement or symbol string that $S=W_1W_2...W_n$ can be calculated by initial probability distribution and transition probability based on the theory of Markov randomized procedure [8]. To establish *N*-gram model, parameters needs to be estimated. The most common way to estimate parameters in the statistical language model is maximum likelihood estimation. The design procedures for text proofreading module for interrogative sentences is as follows:

Step 1: Implement word segmentation for the interrogative sentences submitted by students and tag their part of speech.

Step 2: Build model for words and expressions after word segmentation by adopting *N*-gram language model.

Step 3: Locate incorrect characters for corpus collected by adopting *N*-gram method and detect possibly incorrect words.

Step 4: Correct incorrect character string and obtain the similar words and expressions with greatest probability under *N*-gram model.

Step 5: Calculate the similarity value for words in Step 4 with terminological dictionary. If the similarity value is higher than threshold value that is set for a word, then it is replaced with a correct word in the dictionary and takes the result as a suggested word for students' spelling mistakes, or the result in the previous step is considered as a suggested word.

Step 6: Repeat Step 4 and Step 5 to process each possibly incorrect word.

Step 7: Integrate corrected words and sentences and send back to students as a hint for correcting mistakes in interrogative sentences.

The similarity value between words was calculated using editing distance similarity [9], in which each edition for a character (insert, delete or replace) adds edition distance by one.

IV. Generation of Answers Based on Automatic Summarization for Multiple Documents in Web

A. Extraction of Webpage Texts

The system built in this paper may extract and analyze the texts on the Web page for the questions that are not recorded in FAQ. The diverse webpage sources and different webpage structures require a general extraction method for texts, and we adopted the extraction method for texts proposed by CHEN Xin from Information Retrieval Research Center of Harbin Institute of Technology [10]. The overall idea of this method is that the position of texts is determined by the distribution intensity of Chinese words in the webpages, which is not only simple with efficient webpage analysis but also free from tags in webpages.

B. Generation of Answers Based on Textrank Algorithm

The answer summarizations were generated from multiple documents extracted by adopting TextRank algorithm. TextRank algorithm is to work out a general evaluation standard to give scores for each sentence in texts, which indicate the importance of the sentence, and the sentences in top ranks will be finally acquired and formulated into the abstract of texts [11]. According to TextRank algorithm, the adjacency among internal sentences in a text can be illustrated in a directed graph as "Fig. 2".

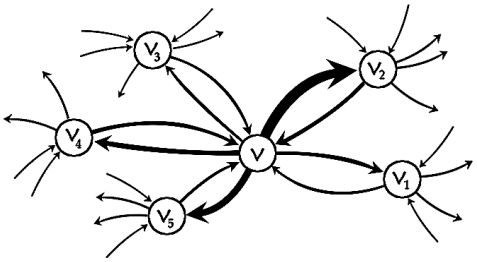

Fig. 2. Graph structure of textrank algorithm.

V. Experiment Results Analysis

Total 896 questions were collected for the course *Fundamentals of Computers*, among which 564 questions were submitted by students and 332 questions were used as the test questions sorted manually, and 411 questions of the test questions were not recorded into FAQ. The tests mainly included text proofreading for interrogative sentences, the quality of answer abstract generated from multiple documents retrieved in Web and overall system operation performance.

Three indicators, with definition in (1), (2) and (3), accuracy rate P, recall rate R and F value were adopted to evaluate test results, where P indicates the accuracy rate of data processing of the QA system, R indicates the capacity of the system to cover correct data, and F is the harmonic mean of accuracy rate P and recall rate R.

$$P = \frac{N_C}{N_R} \times 100\% \tag{1}$$

$$R = \frac{N_C}{N_T} \times 100\% \tag{2}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{3}$$

Wherein, N_R represents the total number of data processed by system such as the number of question-answer couplets identified or the total number of errors detected in question texts; N_C represents the total number of data processed correctly such as the number of answers that meet the requirements in similarity to correct answers (close to questions) in the answers back to students or the total number of errors detected correctly in question texts; N_T represents the total number of data in the test set such as correct number of question-answer couplets or number of errors in question texts.

Correction rate C was used to illustrate system's capability of error correction for the text proofreading performance, which, as defined in (4), is the ratio between the quantity of incorrect texts corrected by system N_{Cr} and the quantity of incorrect texts detected by system N_E.

$$C = \frac{N_{Cr}}{N_E} \times 100\% \tag{4}$$

Based on the analysis on data from test results, total 185 interrogative sentences suffered errors in words and expressions with 218 errors among 896 test questions. Possibly incorrect words and expressions in interrogative sentences were detected through N-gram language model and errors were corrected with terminological dictionary. According to the computational method for evaluation indicators in (1)~(4), the accuracy rate obtained is 84.5%, recall rate is 78.6%, F value is 81.4% and correction rate is 85.3% for text proofreading in interrogative sentences, which indicates that errors in words and expressions can be effectively detected in the interrogative sentences submitted by students.

As currently there is no generally accepted and united standard to evaluate automatic abstracts for Chinese multiple documents, the conformity between the answers automatically generated by system and the answers proposed by teachers were divided into Grade A (high), Grade B (moderate) and Grade C (unacceptable) for evaluation based on the similarity of text semantic information judged manually, and the total number of answers for Grade A and Grade B, as the number of correct answers, were used to calculate accuracy rate. 411 questions retrieved by Web were tested and the quantity of answers for Grade A, Grade B and Grade C were summarized. There are 89 answers with Grade A, 235 ones with Grade B and 87 ones with Grade C, and the accuracy rate is 78.8%, which indicates that the answer abstracts generated from multiple documents that are retrieved by Web is consistent with the answers proposed by teachers to the questions that are not recorded in FAQ.

Total 782 questions with correct answers generated by system were summarized from 896 questions in this paper. And the total number of questions processed by system is 896 and the question-answer couplets with correct answers in the test set are 876. According to the evaluation indicators as shown in (1)-(4), the accuracy rate, recall rate and *F* value of the system is as shown in "Table Ⅰ".

TABLE I. EVALUATION INDICATOR

Total Number of Questions Tested	Accuracy Rate (%)	Recall Rate (%)	*F* Value (%)
896	87.3	89.3	88.3

It can be seen from "Table I", as the results of evaluation indicator, that the *F* value of the system is 88.3%, which further means that the accuracy rate and recall rate of questions are high. Meanwhile, the QA system realized in this paper was compared to other QA systems described in [4], [12], [13] and [14] in accuracy rate, recall rate, and *F* value. The comparison results of evaluation indicators are as shown in "Table II".

TABLE II. EVALUATION INDEX OF DIFFERENT SYSTEMS

Question Answering System for Courses	Accuracy Rate (%)	Recall Rate (%)	*F* Value (%)
Ref. [4]	67.0	-	-
Ref. [12]	92.5	-	-
Ref. [13]	85.1	83.7	84.4
Ref. [14]	84.5	75.7	79.9
This paper	87.3	89.3	88.3

We can see from comprehensive comparison in "Table Ⅱ" that the F value of the system constructed in this paper is higher than those of other QA systems. The system realized in [4] only searches the answers to associated questions in FAQ database with low accuracy rate and few test question sets. The system realized in [12] searches the answers in the existing documents, and there is no solution if answers cannot be found in FAQ database, so the ability to answer questions is limited. And the system realized in [13] is a QA system in a community, which greatest flaw is that questions submitted by students cannot be answered in a timely manner but wait other students to answer. The QA system designed in [14] is a system based on search engine where there is no local FAQ database, in which the answers can only be found when the network is available, and it cannot be used without network.

The QA system constructed in this paper stored regular question-answer couplets for course in database and implemented online search for questions that are not recorded in FAQ, in which incorrect words and expressions in interrogative questions can be corrected and answers can be generated through multiple documents abstract technology when Web is retrieved for answers. We can learn from the F value in Table Ⅱ that the system constructed in this paper is better than other QA systems for courses, in which students' questions can be answered effectively and the system can be used when the network is not available.

VI. CONCLUSION

An intelligent QA system for courses for MOOC environment was constructed in this paper to compensate the deficiencies that answers are available to students' questions only when administrators are online. The system will not only help students to obtain answers of the questions about courses in a timely manner but assist teachers in teaching. The system will examine the rationality of interrogative sentences and correct them effectively. The interrogative sentences submitted by students can be examined through integration of *N*-gram model and dictionary verification, and furthermore, corresponding correction suggestions can be proposed. The accuracy and comprehensiveness of answers was enhanced by calculating and sorting out the importance of sentences in a text through the extraction of multiple webpage texts and multiple documents abstract technology.

A universal, handy and portable question answering mode for teaching in colleges and universities based on MOOC was initially constructed in this paper, which can provide comprehensive theory analysis frame and reference for analyzing, constructing and implementing MOOC teaching mode in colleges and universities. And question answering mode in MOOC teaching can be made correspondingly for specific subject and major based on question answering mode.

REFERENCES

[1] W. Bo, Design and implementation of MOOC with course recommendation, Xidian University master's degree dissertations, 2014, pp. 1-3.

[2] D. Wu, Y. J. Chen, H. Su, H. B. Wang. NMC Horizon Report: 2013 Higher Education Edition (I). Journal of Guangzhou Open University, 2013, pp. 107-112.

[3] W. Wen-li, The development of MOOC and its impact on higher education, Jiangsu Higher Education, vol. 2, 2013, pp. 53-57.

[4] L. Xiang, Research and application of question answering system based on course knowledge, Dalian Maritime University master's degree dissertations, 2010, pp. 1-5.

[5] G. Yi, W. Xiao-long, A statistical measure of semantic similarity between chinese words, The Seventh National Symposium on Computational Linguistics, Harbin, 2003, pp. 221-227.

[6] L. Qun, L. Su-jian, Word similarity computing based on How-net, The Tenth Chinese Lexical Semantics Workshop, Taipei, 2002, pp. 8-15.

[7] X. Bai, Query correction based on N-gram model, Guangxi University master's degree dissertations, 2011, pp. 6-11.

[8] C. Zhi-peng, L. Yu-qin, L. Hua-sheng, L. Gang, T. Hui, Chinese spelling correction in search engines based on N-gram model, Journal of China Academy of Electronics and Information Echnology, vol.4, 2009, pp. 323-326.

[9] L. Liang-liang, C. Cun-gen, Study of automatic proofreading method for non-multi-character word error in Chinese text, Computer Science, vol.43, 2016, pp. 200-205.

[10] L. Zeng-jian, Question answering system based on web search, Harbin Institute of Technology master's degree dissertations, 2013, pp. 13-16.

[11] R. Mihalcea, P. Tarau. TextRank: bringing order into texts. Proc. 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain, 2004, pp. 404-411.

[12] W. Xiu-ping, Research of intelligent answering system based on Chinses word segmentation and ontology, Hunan University master's degree dissertations, 2015, pp. 1-4.

[13] C. Xin-guang, The study and realization of curriculum knowledge based on community question answermg, Chongqing University master's degree dissertations, 2016, pp. 3-7.

[14] Z. Jie, Design and implementation of Chinese automatic question answering system based on search engine, Beijing University of Technology master's degree dissertations, 2016, pp. 7-12.