Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 3167835, 11 pages
https://doi.org/10.1155/2020/3167835

Research Article

Research on Chinese Question-Answering for Gaokao Based on Graph

Zhizhuo Yang,[1] **Chunzhuan Li**,[1] **Zhang Hu**,[1] **Qian Yili**,[1] **and Ru Li**[1,2]

[1]*School of Computer and Information Technology of Shanxi University, Taiyuan, Shanxi 030006, China*
[2]*Key Laboratory of Computation Intelligence and Chinese Information Processing, Taiyuan, Shanxi 030006, China*

Correspondence should be addressed to Zhizhuo Yang; yangzhizhuo@sxu.edu.cn

Received 17 August 2020; Revised 24 October 2020; Accepted 26 October 2020; Published 6 November 2020

Academic Editor: Jun Shen

Reading comprehension Question-Answering (QA) for College Entrance Examination (Gaokao in Chinese) is a challenging AI task because it requires effective representation to capture complicated semantic relations between the question and answers. In this paper, a novel method of Chinese Automatic Question-Answering based on a graph is proposed. The method first uses the Chinese FrameNet and discourse topic (paragraph topic sentence and author's opinion sentence) to construct the affinity matrix between the question and candidate sentences and then employs the algorithm based on the graph to iteratively calculate the importance of each sentence. At last, the top 6 candidate answer sentences are selected based on the ranking scores. The recall on Beijing College Entrance Examination in the recent twelve years is 67.86%, which verifies the effectiveness of the method.

1. Introduction

Teaching the computer to pass the entrance examination of different education levels, which is an increasingly popular artificial intelligence challenge, has been taken up by researchers in several countries in recent years [1–3]. The Todai Robot Project [3] aims to develop a problem-solving system that can pass the University of Tokyo's entrance examination. China has launched a similar project "key technology and system for language question solving and answer generation," focusing on studying the human-like QA system for College Entrance Examination (commonly known as Gaokao). Gaokao is a national-wide standard examination for all senior middle school students in China and has been known for its large scale and strictness.

Although deep learning methods have achieved good results in many natural language processing tasks [4–7], they usually rely on a large scale of the dataset for effective training. However, the Gaokao task cannot receive sufficient training data under the current conditions. Different from previous typical QA tasks such as SQuAD [8], DuReader [9], and CMRC2018 [10] which can enjoy the advantage of holding a very large known QA pair set, the concerned task is equal to retrieving a proper answer from background article with guidelines of a very limited number of known QA pairs. In addition, the questions are usually given in an implicit way to ask students to dig the exactly expected meaning of the concerned facts. If such kind of meaning fails to fall into the feature representation for either question or answer, the retrieval will hardly be successful.

Generally speaking, for the Gaokao challenge, knowledge sources are extensive and no sufficient structured dataset is available, while the most existing work on knowledge representation focused on structured and semistructured types [11–14]. With regard to the answer retrieval, there are models based on semantic resources such as HowNet [15], WordNet [16], and Synonym Cilin [17]. Reference [18] proposed a sentence semantic relevance calculation method based on the multidimensional voting algorithm. This method considers the semantic relevance of different dimensions as a metric and uses the idea of the voting algorithm to select the best option for the problem. Reference [19] proposed a title selection method based on a correlation matrix between the title and the main points of the chapter. Reference [20] proposed a method for extracting candidate sentences based on frame matching and

frame relationship matching and then used manifold ranking to sort the candidate sentences.

This work focuses on reading comprehension question-answering in Gaokao Chinese examinations, which accounts for a large proportion of total scoring and is extremely difficult in the exams. Reference [2] made a preliminary attempt to take up the Gaokao challenge and proposed a three-stage approach that exploits and extends information retrieval techniques. Differently, this task is to solve reading comprehension questions and has to be based on deep semantic representation and computation rather than word matching in the previous work. Table 1 shows an example question in Chinese exams, consisting of a question and answer to the question. Some answer sentences are difficult to retrieve through literal matching, and these answer sentences are not distributed in a paragraph, but in different paragraphs of different articles. For instance, the question sentence would be confusing without knowing about the background article making cultural relics "live." In addition, some answers summarize the article from different paragraphs, while other answers summarize the author's point of view. How to retrieve those answers hidden in scattered paragraphs is a large challenge, and it is also the key to improving the effect of the system for Gaokao.

The challenge of our task would call for a new problem-solving framework for automatically answering comprehensive questions in exams. We propose a graph-based framework as shown in Figure 1. Firstly, we preprocess the articles and questions, and the evidence is drawn. Secondly, the Chinese FrameNet and discourse topic are used to construct the affinity matrix, which preserves the results of the semantic analysis of the question and each sentence. Finally, reasoning is performed by a graph-based ranking algorithm to check each candidate sentence, and the most relevent candidate sentence to the question will be returned as the answer.

Our contribution is threefold: (1) after showing Gaokao's difficulty and its difference from the existing research problems, we propose a new framework for reading comprehension QA in Gaokao. It is the first time to apply a graph-based algorithm in reading comprehension QA. (2) To the best of our knowledge, the relationship between candidate sentences has not been taken into account in the QA task. The relationship between candidate sentences is considered as a factor in our method, and the answer sentences are extracted by the unified model to improve the answering effect of the QA system. (3) Our approach achieves encouraging results on a set of real-life questions collected from recent Chinese examinations. We also release a Chinese comprehensive deep question-answering dataset to facilitate the research.

2. Reading Comprehension QA Method Based on Graph

2.1. Method Framework. The graph-based model [21] was firstly used by search engines to calculate the importance of webpages. It has been successfully used in many tasks, such as object retrieval [22], keyword extraction [23], and automatic summarization [24]. The algorithm is based on the following two assumptions. (1) Quantity assumption: in the web graph model, if a web page A is linked by a lot of other webpages, then page A is more important. (2) Quality assumption: if a page node A is linked by other higher-quality pages, then the A page is more important. The reading comprehension QA graph proposed in this paper is derived from the PageRank model. This model makes full use of the correlation between the question and candidate sentences. The global optimization ranking model is used to extract and sort the answer candidate sentences. The model is based on the following three hypotheses. (1) Quantity hypothesis: if an answer candidate sentence is associated with more other sentences, then the answer candidate sentence is more likely to be an answer sentence. (2) Quality hypothesis: if an answer candidate sentence is associated with other sentences of higher quality, then the answer candidate sentence is more likely to be an answer sentence. (3) Link weight hypothesis: the higher the degree of correlation between the question and the answer candidate sentence is, the more likely the answer candidate sentence is the answer sentence.

This paper makes use of the"voting" or "recommendations" between the question and sentences in the QA problem. The graph for reading comprehension QA is shown in Figure 2. The squares represent the candidate sentences $\{S_1, S_2, \ldots, S_n\}$ in the background article, and the edges between the squares represent the relationship between the candidate sentences, which is represented by the affinity matrix W_{ij}. The upper round node represents the question S_0. Usually, the College Entrance Examination has 1 or 2 questions. If there are 2 questions, they are merged into 1 sentence. The dotted line indicates the relationship between the question and candidate sentence nodes and is represented by the relationship matrix W_{0i} or W_{i0}. In the graph, the initial value of S_0 is set to 1, and the initial value of other candidate sentence nodes is 0. The importance of S_0 is passed to the candidate sentence node through the matrix W_{0i}. At the same time, the importance of the candidate sentences will also be strengthened with each other through the matrix W_{ij}. The importance of the candidate sentence nodes converges to a fixed set of values, and then the candidate sentence nodes are sorted according to the importance score. Finally, the top 6 sentences are selected as the final answer sentences. The difference between reading comprehension QA graph and PageRank graph is that, in PageRank network graph, the type of edge connecting nodes is the same, which indicates the recommendation of two website nodes; while the type of edge of reading comprehension graph is different, one is the edge between question and candidate sentence, which represents an association of answer or explanation. The other is the edge between candidate sentence nodes, which represents an association of similar contents between candidate sentences.

In this paper, the function $f: X \longrightarrow R$ is defined as a ranking function, which assigns a ranking score value f_i to each node S_i. f can be seen as a vector $f = [f_0, f_1, \ldots, f_n]^T$. The definition vector $y = [y_0, y_1, \ldots, y_n]^T$ represents the initial value of each node, where $y_0 = 1$, and the remaining $y_i = 0$. The algorithm is as shown follows:

Table 1: Example of reading comprehension QA in College Entrance Examination.

2017 Beijing College Entrance Examination question
Question: 请结合上述三则材料, 简述让文物“活”起来的含义与作用 Please combine the above three materials to briefly describe the meaning and function of making cultural relics “live.” 答案, 利用博物馆、各种现代技术让参观者近距离感悟文物的魅力。发挥它们在公众知史爱国, 鉴物审美, 以及技艺传承、文化养心的作用, 实现学术、趣味性统一, 以新鲜时尚的方式提供给观众审美与求知、娱乐与鉴赏的多元文化体验, 借助计算机等生成三维环境, 调动多感官, 带来沉浸感, 使用现代技术使得文物呈现方式灵活, 让更多的人喜欢上古文化, 更好地实现文物走近大众的作用。解决了展出空间有限、文物损毁等问题, 起到更好地保护文物的作用 Answer: use museums and various modern technologies to make visitors feel the charm of cultural relics up close. Play their role in public knowledge of history, patriotism, appreciation of objects, as well as technical inheritance, and cultural cultivation; achieve the unity of academic and interesting; provide audiences with a multicultural experience of aesthetics and knowledge, entertainment, and appreciation in a fresh and fashionable way; and use computers to generate a three-dimensional environment, mobilize multiple senses, and bring immersion; the use of modern technology makes the presentation of cultural relics flexible, so that more people like ancient culture, and better realize the role of cultural relics reaching the public. **(paragraph topic sentence)** It solves the problems of limited exhibition space and damage to cultural relics and plays a better role in protecting cultural relics. **(author's opinion sentence)**

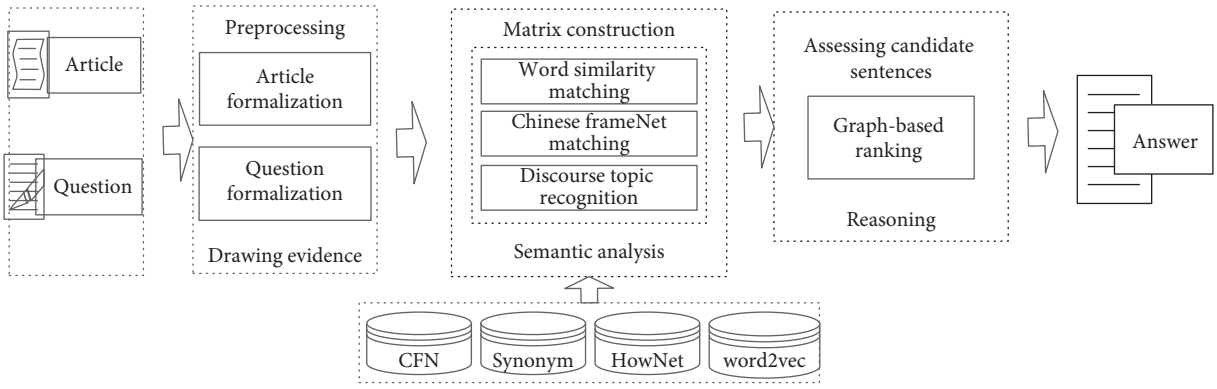

Figure 1: Overview of the approach.

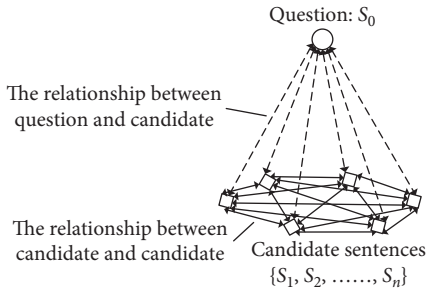

Figure 2: Reading comprehension QA graph for Gaokao.

In the first step of the algorithm, the relationship between the question S_0 and each candidate sentence $\{S_1, S_2, \ldots, S_n\}$ is calculated by the method based on word similarity matching, frame matching, and discourse topic. How to measure the relationship between the question and candidate sentences is the key step of the automatic QA method. For details, see Section 2.2.

In the second step of the algorithm, since the task of this paper is automatic QA, and the answer candidate sentences need to be extracted. The importance transmitted between candidate sentences should be related to the question, and the importance not related to the question should not be transmitted to each other. Therefore, the following formula is used to calculate the relationship between candidate sentences:

$$W_{ij} = \frac{\left(e_{i0} + e_{j0}\right)}{2}, \tag{1}$$

here $i, j \in [1, n]$ and e_{i0} and e_{0j} represent the similarity between the candidate sentences S_i and the question sentence S_0, respectively. The similarity of sentences is calculated by formula (5).

In the third step of the algorithm, the high-quality answer sentences are all explanations and answers to the question. The extraction effect depends largely on the relationship between the candidate sentences and the question, and it is less affected by the relationship between the candidate sentences. Therefore, different weights should be set for the affinity matrix of the two parts. $(\eta_1 > \eta_2)$ means that the relationship between the question and the candidate answer plays a larger role, and the relationship between the candidate sentences plays a smaller role. Previous studies have only focused on the relationship between the question sentence and answer sentences while ignoring the relationship between candidate sentences, but we believe that introducing the relationship between candidate sentences can improve the effect of the QA system. For example, a candidate sentence S_i is not only related to the question sentence but also related to other

Input: question S_0 and answer set $\{S_1, S_2, \ldots, S_n\}$, sentences initial value vector y.
Output: top 6 answer candidate sentences.

(1) Calculate the relationship between the question S_0 and each candidate sentence $\{S_1, S_2, \ldots, S_n\}$ by methods based on word similarity matching, frame matching, and discourse topic. If the degree of relationship between two nodes is greater than 0, the nodes are connected by an edge. Construct the affinity matrix $W_{0i} = W_{i0} = \text{relation}(S_0, S_i)$. In order to prevent the self-reinforcement of each node, let $W_{ii} = 0$.
(2) Calculate the relationship between each candidate sentence through the word similarity. If the degree of relationship between two nodes is greater than 0, the nodes are connected by an edge. Construct the affinity matrix W_{ij}, while $W_{ii} = 0$.
(3) Combine the affinity matrix and normalize it. Define $W = \eta_1(W_{0i} + W_{i0}) + \eta_2 W_{ij}$. Define the diagonal matrix D, where D_{ii} represents the sum of the i-th row of the W, and the W is normalized to $S = D^{(1/2)} W D^{-(1/2)}$.
(4) Iterate $f(t+1) = \alpha S f(t) + (1-\alpha) y$ until convergence, where $\alpha \in [0, 1]$.
(5) Use f_i^* to represent the convergence sequence $\{f_{i(t)}\}$, so that each sentence gets its ranking score.
Return top 6 candidate sentences with the highest score.

Algorithm 1: QA algorithm for reading comprehension based on a graph.

candidate sentences in the background article; then this candidate sentence S_i can represent other candidate sentences to a certain degree, so the candidate sentence is more likely to be the answer sentence. In this step, the affinity matrix is normalized to ensure the convergence of the iterative algorithm.

In the fourth step of the algorithm, α is a key parameter of the graph-based algorithm. This parameter can balance the impact of neighboring nodes and the initial scores of other nodes: the closer α is to 1, the greater the influence of neighboring nodes on the score; the closer α is to 0, the greater the influence of the initial score of nodes on the score. When the affinity matrix S satisfies the Markov process convergence conditions, the importance of the nodes converges. Usually, the convergence of the iteration algorithm is achieved when the difference between the scores computed at two successive iterations for any point falls below a given threshold (0.0001 in this paper).

By counting the suggested answers of the examination papers over several years, it is found that the average number of answer sentences is 6. If the number of outputs is less than 6 sentences, it is not enough to cover all answer points; if the number of outputs is greater than 6 sentences, the redundancy of the output answers is high. Finally, the top 6 candidate sentences are selected as answer sentences by the algorithm.

2.2. Calculation of the Relationship between the Question and Candidate Sentences. The calculation of the relationship between the question S_0 and each candidate sentence $\{S_1, S_2, \ldots, S_n\}$ directly affects the final answer. This paper uses a novel method based on CFN [20] and discourse topic to calculate the relationship of the question and candidate sentences. Our method takes into account sentence similarity, sentence frame matching, and discourse topic matching. The affinity matrix is used to record the relationship between the question and candidate sentences. The affinity matrix is a symmetric matrix. The calculation formula is as follows:

$$W_{0,i} = W_{i,0} = \lambda_1 * W_1 + \lambda_2 * W_2 + \lambda_3 * W_3 + \lambda_4 * W_4, \tag{2}$$

where $i \in [1, n]$, W_1 represents the sentence similarity matrix, W_2 represents the sentence frame matching matrix, W_3 represents the paragraph topic sentence matrix, W_4 represents the author's opinion sentence matrix, λ_k is the weight of the k-th dimension, $k \in [1, 4]$, and $0 \le \lambda_k \le 1$, $\sum_{k=1}^{4} \lambda_k = 1$. λ_k is used to adjust the weight of each matrix, and the value of the weight is set in the experiment.

2.2.1. Answer Sentence Extraction Based on Similarity Measure. First, preprocess the sentence, including word segmentation and removal of stop words. $S_0 = \langle k_1, k_2, \ldots, k_m \rangle$, $S_i = \langle w_1, w_2, \ldots, w_m \rangle$, and k_i and w_j represent the keywords of the question and candidate sentences, respectively; then, we combine HowNet [11] and word2vec [25] to calculate the similarity as follows:

$$\begin{aligned} \text{sum word} = {} & 0.4 \times \max_{1 \le i,j \le n} \left(\text{simHownet}\left(K_i, W_j\right)\right) \\ & + 0.6 \times \text{Cos}\left(K_i^v, W_j^v\right), \end{aligned} \tag{3}$$

where $\text{simHownet}(k_i, w_j)$ means calculating the similarity between the keyword k_i and w_j by HowNet. We use word2vec to calculate the cosine similarity of a word vector as follows:

$$\text{Cos}\left(k_i^v w_j^v\right) = \frac{k_i^v \cdot w_j^v}{\left(\left\|k_i^v\right\| \times \left\|w_j^v\right\|\right)}, \tag{4}$$

where k_i^v and w_j^v represent the word vectors of k_i and w_j. Finally, normalize (sumword_i) and the final calculation formula is

$$\begin{aligned} W_1 = W_{0i} = W_{i0} &= \text{Score}_{\text{sumWord}} \\ &= \frac{\text{sumword}_i}{\left\{\max_{1 \le i \le n}(\text{sumword}_i) - \min_{1 \le i \le n}(\text{sumword}_i)\right\}}. \end{aligned} \tag{5}$$

2.2.2. Answer Sentence Extraction Based on Frame Matching. Since the method based on similarity measure cannot mine the deep semantic information of the sentences in Gaokao, this paper uses the Chinese Frame Network (CFN) [26] to capture the semantic information in the semantic scene.

CFN is a Chinese vocabulary semantic knowledge base established by Shanxi University; it is based on FrameNet [27] of the University of California, Berkeley.

(1) Frame semantic matching: when the frame evoked by the target word of the question S_0 is the same frame evoked by the target word of the sentence S_i, the matching number is increased by one. (2) Frame semantic relationship matching: when the distance between the frame evoked by the question S_0 and the frame evoked by the sentence S_i is less than or equal to 2, the matching number is increased by one. Then, the frame matching number of the candidate sentence and the question is obtained. Finally, normalize it and the score based on frame matching is

$$W_2 = W_{0i} = W_{i0} = \text{Score}_{\text{sumFrame}} = \frac{\text{sumframe}_i}{\{\max_{1\le i\le n}(\text{sumframe}_i) - \min_{1\le i\le n}(\text{sumframe}_i)\}}. \tag{6}$$

An example of candidate sentence extraction based on frame matching is shown in Figure 3. The frame aroused by the target word "*development*" in question is the same as that aroused by the target word "*enhance*" in the candidate sentence; there is a relationship between the frame aroused by the target word "*development*" in the question and the frame aroused by the target word "*carry out*" in the candidate sentence. The involved scenes are relevant and the distance is less than or equal to 2. Therefore, the sentence S is extracted as an answer candidate sentence based on frame matching.

2.2.3. Answer Sentence Extraction Based on Discourse Topic. Through the study of the examination outline, it is found that College Entrance Examination often inspects the ability of students to summarize the main idea of the article. This paper proposes a method of extracting candidate sentences based on the discourse topic, which includes paragraph topic sentences and author opinion sentences.

2.2.4. Paragraph Topic Sentence Extraction. Through researching a large number of examination papers, it is found that the topic sentences are usually located at the beginning or end of the paragraph, and the sentence is usually related to the topic of other sentences in this paragraph. As shown in Table 1, "Use computers to generate a three-dimensional environment, mobilize multiple senses, and bring immersion," is located at the beginning of the paragraph and it is a topic sentence in the paragraph.

(1) Position Information. Paragraph topic sentence is a summary of the paragraph, which reflects the main idea of the paragraph. It is generally distributed at the beginning or end of the paragraph. Therefore, each sentence is calculated according to the position of the paragraph:

$$\text{score}_i = \begin{cases} 1, & i = 1, n, \\ 1 - \dfrac{\log i}{\log n}, & \text{others}, \end{cases} \tag{7}$$

where i is the sentence number and n is the total number of sentences in each paragraph.

For different paragraphs, in general, the first and last paragraphs of the article can reflect the topic of the article, so the weight of the first and last paragraphs should be greater, and the topic sentence of each paragraph is calculated according to the position of the paragraph:

$$\text{score}_{\text{loc}} = \begin{cases} 0.7 \times \text{score}_i, & i = 1 \text{ or } i = m, \\ 0.3 \times \text{score}_i, & \text{others}, \end{cases} \tag{8}$$

where m is the total number of paragraphs in the article.

(2) Semantic Similarity between Sentences Based on Paragraph. The keyword of sentence A is A_i, with p in total, and the keyword of sentence B is B_j, with q in total.

HowNet is used to calculate the similarity of sentences. The similarity of two words based on HowNet is $S(A_i, B_j)$. Let $a_i = \max\{S(A_i, B_1), S(A_i, B_2), \ldots, S(A_i, B_p)\}$, $b_j = \max\{S(B_j, A_1), S(B_j, A_2), \ldots, S(B_j, A_q)\}$; then, the similarity of sentences based on HowNet [11] is

$$\text{sim}(A, B) = \frac{\left(\sum_{i=1}^{p} a_i/p\right) + \left(\sum_{j=1}^{q} b_j/q\right)}{2}. \tag{9}$$

Then, the semantic similarity of sentence A based on paragraphs is

$$\text{score}_{\text{sim}} = \frac{\sum_{x=1}^{n} \text{sim}(A, B_x)}{n}, \tag{10}$$

where n is the total number of sentences in each paragraph.

Finally, the above factors are weighted to obtain the calculation formula as follows:

$$W_3 = W_{0i} = W_{i0} = \text{Score}_{\text{topic}} = \beta_1 * \text{score}_{\text{loc}} + \beta_2 * \text{score}_{\text{sim}}, \tag{11}$$

where $\beta_1 + \beta_2 = 1$.

2.2.5. Author's Opinion Sentence Extraction. It is found that the author's opinions and attitudes often appear in the suggested answers. The opinion sentences mainly indicate the author's viewpoint and attitude in the article, which are the overall grasp of the content and the topic of the whole discourse. Position information, similarity between sentences, and suggestive words are important features of author opinion sentences. As shown in Example 1, sentence S is the first sentence of the last paragraph in the article, and, secondly, the sentence is semantically related to other sentences, indicating the author's attitude in the whole article.

Example 1. 2018 Beijing College Entrance Examination.

Question: 根据材料一、材料二, 简要说明人类对人工智能的认识是如何不断深化的。

According to material one and material two, briefly explain how humans have continuously deepened their understanding of artificial intelligence.

Sentence: 面对人工智能可能带来的种种冲击, 上世纪50年代美国科幻小说家阿西莫夫提出的机器人三大定律, 今天对我们依然有借鉴意义。

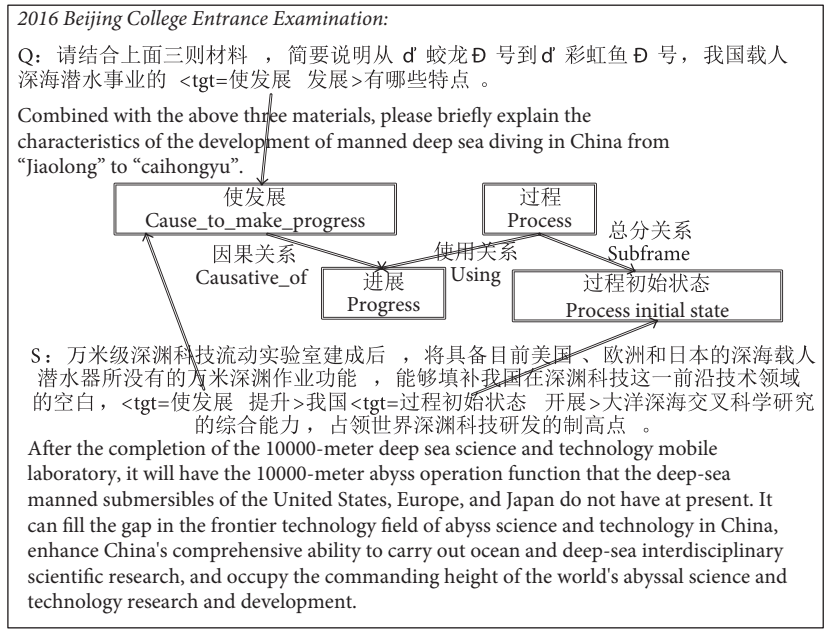

Figure 3: An example of candidate sentence extraction based on frame matching.

Faced with the various impacts that artificial intelligence may bring, the three laws of robotics proposed by the American science fiction novelist Asimov in the 1950s still have reference significance for us today.

(1) Position information

By analyzing the examination papers, it is found that the author's point of view is generally distributed at the end of the article, and the calculation is based on the different positions of the sentence in the last paragraph. The calculation formula is as formula (7), which is recorded as (score_i).

(2) Semantic similarity between sentences based on paragraph

The semantic similarity between sentences is calculated when extracting the author's opinion sentences. The calculation formula is as formula (10).

(3) Heuristic rules based on suggestive words

Candidate sentences are extracted based on whether the sentence contains suggestive words. If the sentence contains suggestive words, $\text{score}_{\text{Word}} = 1$; otherwise, $\text{score}_{\text{Word}} = 0$. This article expands the suggestive vocabulary through the CILIN [28]. Examples of suggestive words are shown in Table 2.

Finally, the above three factors are weighted to obtain the score of the author's opinion sentence:

$$\begin{aligned} W_4 &= W_{0i} = W_{i0} = \text{Score}_{\text{opinion}} \\ &= \gamma_1 * \text{score}_i + \gamma_2 * \text{score}_{\text{sim}} + \gamma_3 * \text{score}_{\text{Word}}, \end{aligned} \tag{12}$$

where $\gamma_1 + \gamma_2 + \gamma_3 = 1$.

3. Experiment and Result Analysis

3.1. Experimental Data. In the experiment, the language technology platform LTP [29] was used for word segmentation and part-of-speech tagging. The CFN [26] provided by Shanxi University was used for frame matching, and the HowNet [11] platform was used for word similarity calculation.

Due to the small proportion of questions in the College Entrance Examination, the dataset used in this paper includes the College Entrance Examination in each province, the simulation examination questions, and the questions transformed from multiple-choice questions. Finally, 132 questions were collected on the College Entrance Examination test in each province for the past 12 years, and 511 questions were collected on the simulation examination questions in each province. Each question consists of 1 or 2 questions. On average, each material contains 40 sentences and each sentence contains 30 Chinese words.

3.2. Experimental Results and Analysis

3.2.1. Comparison of Results of Different Experimental Methods. At present, the answers to the examination are graded according to the key points of the questions. In this paper, recall and accuracy are used as evaluation. A fivefold experiment was used to divide the corpus into five parts, one of which was used as the test set and the other four as the training set. The experiment was repeated five times, and the average value was taken as the final result. We manually find several answer sentences from the article according to the suggested answers, and mark them as the set A. S_A is the set of the top 6 sentences sorted by our method:

$$\begin{aligned} R &= \frac{\text{total sentences of correct sentences in } S_A}{\text{total sentences of } A} \times 100\%, \\ P &= \frac{\text{total sentences of correct sentences in } S_A}{\text{total sentences of } S_A} \times 100\%. \end{aligned} \tag{13}$$

Table 2: Examples of suggestive words.

看来、由此可见、由此看来、可见、无论如何、不管怎样、综上所述、由上述可知、如上所述、总的来看、总的来说、总之、总而言之、总体而言、首先、其次、表明、所以
It seems, thus it can be seen, it can be seen, anyway, no matter what, in summary, from the above, as mentioned above, in general, in general, in short, all in all, in general, first, second, show, so

To verify the effectiveness of the method in this paper, the method in this paper is compared with multiple baseline methods on Beijing College Entrance Examination questions for the past 12 years. The baseline methods include the following:

(1) Use frame matching [20] as baseline 1.

(2) We use the BERT model [30] as baseline 2. The model classifies answer candidates into two categories; in other words, it judges whether the candidate sentences in the article are answer sentences. College Entrance Examinations in all provinces except for Beijing and simulation examination questions (including 122 College Entrance Examination questions and 511 simulation examination questions, and we manually mark the answer sentences in the article according to the suggested answers) were used to train the model.

(3) The direct ranking method is used as baseline 3: the scores of each sentence are calculated by linear interpolation of word similarity matching, frame matching, and discourse topic, and the formula is as follows:

$$S = \phi_1 * \text{Score}_{\text{sumWord}} + \phi_2 * \text{Score}_{\text{sumFrame}} + \phi_3 * \text{Score}_{\text{topic}} + \phi_4 * \text{Score}_{\text{opinion}}, \quad (14)$$

where ϕ_k is the weight of the k-th dimension, $k \in [1, K]$, and $0 \le \phi_k \le 1$, $\sum_{k=1}^{K} \phi_k = 1$. In the experiment, we set $\phi_1 = 0.3, \phi_2 = 0.2, \phi_3 = 0.3, \phi_4 = 0.2$.

The experimental results are shown in Table 3.

There are many parameters in the method proposed in this paper, and these parameters are all based on experimental tests. Specifically, fix other parameters, take a parameter value from 0.0 to 1.0 in steps of 0.1, and test it, respectively. When the answer effect is the best, the parameter value is the final value. In Algorithm 1, $\eta_1 = 1.0$, $\eta_2 = 0.1$, $\alpha = 0.6$; when calculating the relationship between the question and candidate sentences, λ_k is set to 0.4, 0.2, 0.2, and 0.2. In the method of extracting the answer sentences based on discourse topic, $\beta_1 = 0.7$, $\beta_2 = 0.3$, $\gamma_1 = 0.3$, $\gamma_2 = 0.1$, and $\gamma_3 = 0.6$.

To compare with the international popular methods in reading comprehension for QA tasks, our method is compared with the deep learning method. It can be found that the recall of the BERT model is only 39.50%, which shows that the application of BERT in the College Entrance Examination is not good. The College Entrance Examination questions are more difficult than ordinary reading comprehension questions, and in the current scale of training data, we cannot train an efficient model which can capture complicated semantic relations between the question and answer. Moreover, the structure of the BERT model is very complex, and it is not easy to add rich linguistic knowledge to the model, which makes the model unable to adapt to the task in specific field.

When the direct ranking method is used, the recall and accuracy are 63.69% and 50.00%, respectively. When the QA method based on the graph is adopted, the recall and accuracy have been further improved. It should be noted that these two methods use the same external knowledge, but different algorithms to extract candidate sentences. The direct ranking method calculates the scores of each candidate sentence on each dimension and then performs a weighted sum of the scores of each dimension; the method based on the graph calculates the scores of each candidate sentence iteratively. The importance of the candidate sentences is transferred in the graph until it is finally stable. The experimental results show that our method can calculate the importance of each candidate sentence more reasonably and accurately.

3.2.2. Comparison of Results of Direct Ranking Method. To prove the advantages of the graph-based method, we use different methods in Section 2.2 to establish the affinity matrix of the question and candidate sentences and then use different methods to perform ranking. The experiment was carried out in the last 12 years of College Entrance Examination in Beijing, and the results are shown in Table 4.

It can be seen from Table 4 that when PageRank ranking is adopted, the experimental results are improved compared to the direct ranking method. The experimental results show the effectiveness of the iterative ranking method. Our method considers not only the relationship between the question and candidate sentences but also the relationship between candidate sentences. The algorithm based on graph can extract candidate sentences with both high degree of relevance to the question and high similarity with other candidate sentences. In addition, the experimental results also show that when four different methods (sord similarity + frame matching + paragraph topic sentence + author's opinion sentence) are used at the same time to extract candidate sentences, the experimental results have reached the optimal value whether it is direct ranking or PageRank ranking. It shows that various methods can make up for each other and jointly improve the effect of the system. λ_k of our method is set to 0.4, 0.2, 0.2, and 0.2, indicating that the word similarity method plays a greater role, while other methods play a smaller role. The last three methods can extract some answers that are not literally similar.

3.2.3. The Effect of $\eta_1 : \eta_2$ on the Experimental Results. $\eta_1 : \eta_2$ indicates the proportion of the relationship between candidate sentences and the question and the relationship between candidate sentences. The experiment was carried out on Beijing College Entrance Examinations, College

Table 3: Comparison results of different methods.

Method	R (%)	P (%)
Baseline 1 (frame matching)	50.48	35.00
Baseline 2 (BERT)	39.50	35.30
Baseline 3 (direct ranking method)	63.69	50.00
Automatic QA method based on graph	67.86	51.67

Table 4: The experimental results are compared with the direct ranking method.

Method		R (%)	P (%)
Word similarity	Direct ranking	48.57	33.33
	PageRank ranking	58.69	43.33
Paragraph topic sentence	Direct ranking	39.41	33.33
	PageRank ranking	50.48	36.67
Author's opinion sentence	Direct ranking	38.45	31.67
	PageRank ranking	51.79	38.33
Word similarity + frame matching	Direct ranking	52.74	36.67
	PageRank ranking	60.60	45.00
Word similarity + frame matching + paragraph topic sentence + author's opinion sentence	Direct ranking	63.69	50.00
	Automatic QA method based on graph	67.86	51.67

Entrance Examinations in other provinces, and simulation examination questions. The experiment fixed $\alpha = 0.6$. The results are shown in Figure 4. It can be found that when $\eta_1 : \eta_2 = 1 : 0.1$, the effect is the best. It proves that the relationship between candidate sentences is beneficial to QA in the College Entrance Examination, and the relationship between candidate sentences plays an auxiliary role, so η_1 is set larger and η_2 is set smaller.

3.2.4. The Effect of α on the Experimental Results. Experiments were carried out on Beijing College Entrance Examination questions, College Entrance Examination questions in other provinces, and simulated examination questions. $\eta_1 : \eta_2 = 1 : 0.1$ was fixed in the experiment. The results are shown in Figure 5. It can be found that the value of α has little effect on Beijing College Entrance Examination questions, while the best results are obtained when $\alpha = 0.6$ on College Entrance Examination questions and simulated questions in other provinces. The experiments show that neighboring nodes have a greater influence on candidate sentence scores, and initial score nodes have less influence on candidate sentence scores.

3.2.5. Differences between Real Questions in Different Provinces and Simulated Questions. It can be seen from Figures 4 and 5 that the method proposed in this paper has the best effect on Beijing College Entrance Examinations, but slightly worse on College Entrance Examinations in other provinces and simulation examinations. Because there are differences between them: the articles in Beijing College entrance examination are usually scientific and technological papers, while the articles in other provinces are mostly papers, academic papers, current reviews, book reviews, news, biographies, reports, popular science, and so on. In addition, most of the questions in Beijing College Entrance Examination are examined to select and integrate the information in the article, while most of the questions in other provinces are examined to understand the important words and sentences and grasp the structure and overall idea of the article. The recall of real and simulated questions in different provinces is shown in Figure 6.

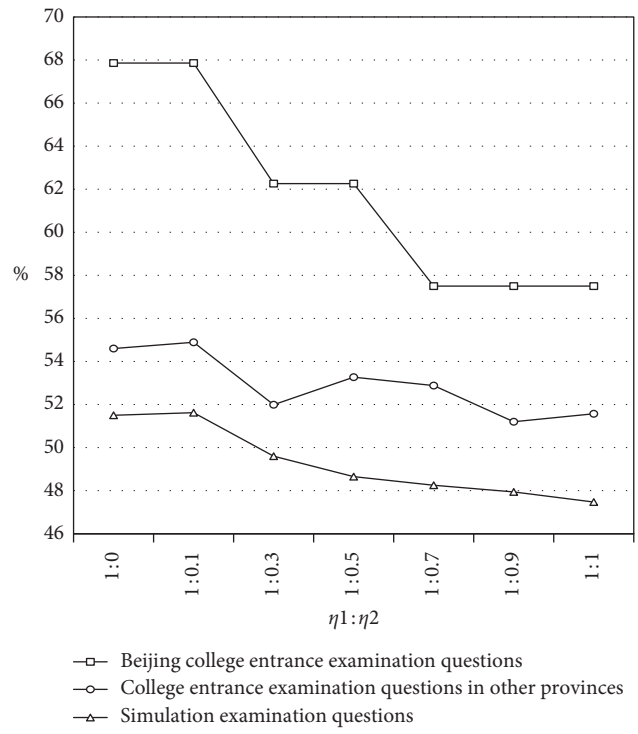

Figure 4: The effect of $\eta_1 : \eta_2$ on the experimental results.

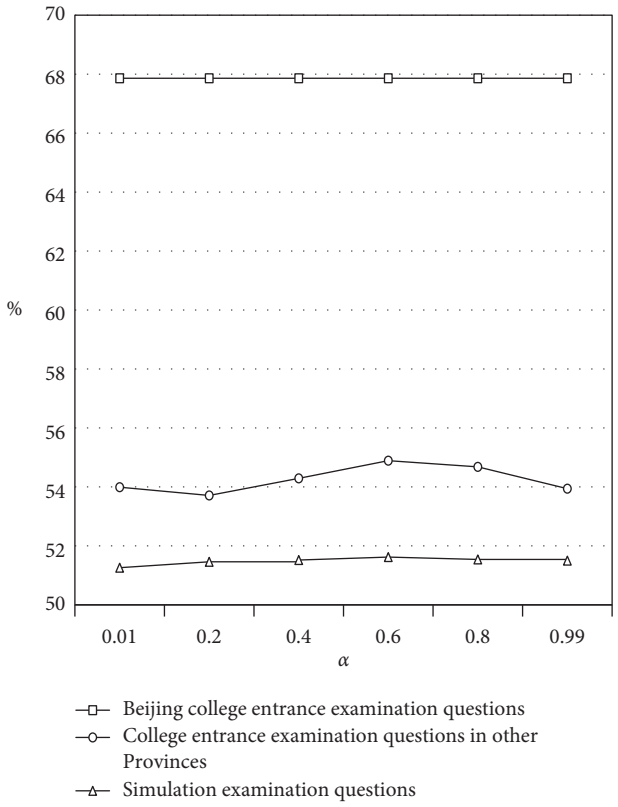

Figure 5: The effect of α on the experimental results.

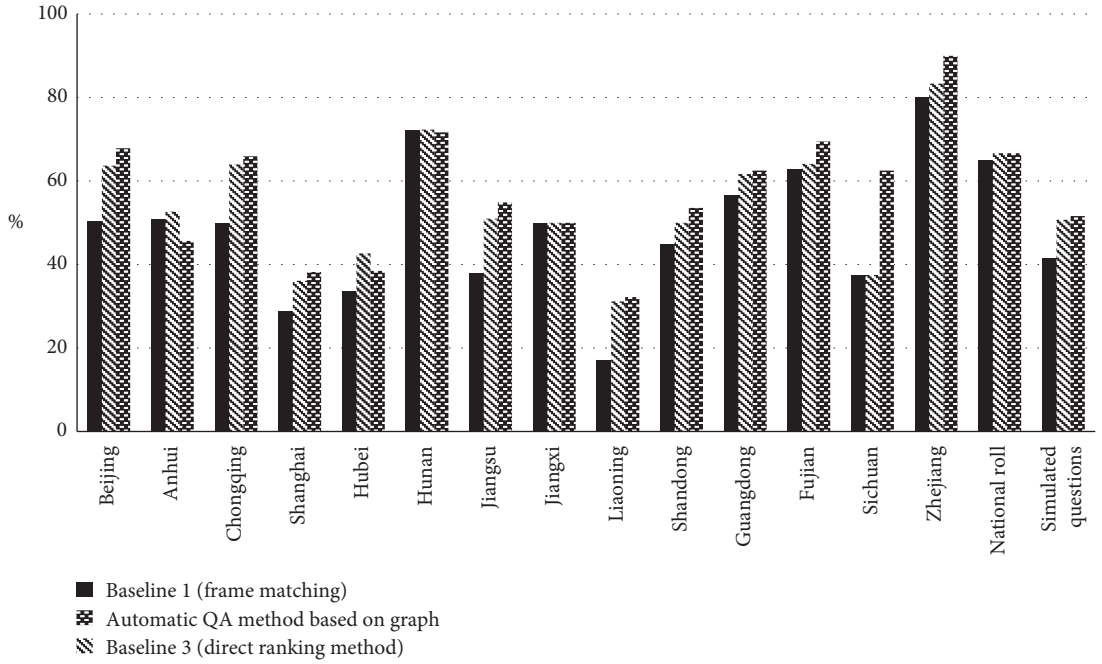

Figure 6: Recall rates of real and simulated questions in different provinces.

It can be seen from Figure 6 that our method can improve the experimental effect on real and simulated questions in different provinces. At the same time, it can be found that the recall of some provinces is relatively low, such as Example 2.

Example 2. 2009 Liaoning College Entrance Examination questions.

Question: "通俗历史热"在当今出现的原因是什么?

What is the reason for the emergence of "popular history fever" today?

Answer: "通俗历史热"是商品经济和文化教育发展到一定程度后定会出现的一种现象。

当商品经济趋于发达、文化教育发展迅速的时候, 人们在从事赖以谋生的职业活动之外, 带有文化色彩的业余需求会随之增长, 对作为文化存在常见形态之一的历史知识, 其"求解"欲望也会趋于强烈。

在当今市场经济逐步成熟、文化教育普及程度大为提高、高等教育开始走向大众化的时代, 人们的业余文化需求显著增长, 久远的尘封旧事引起了人们日益浓厚的兴趣。

对于广大民众而言, 在古奥难懂的传统史著和"学术模式"的现代史书皆难"卒读"的情况下, 通俗化的历史几乎成为他们"探寻过去"的唯一选择。

"Popular history fever" is a phenomenon that will surely appear after the development of the commodity economy and cultural education to a certain extent.

When the commodity economy tends to develop and cultural education develops rapidly, in addition to the professional activities that people rely on to make a living, the demand for culturally colored amateurs will increase accordingly. For historical knowledge as one of the common forms of cultural existence, its desire to "solve" will also become stronger.

In today's era, when the market economy is gradually maturing, the popularity of cultural education has greatly increased, higher education has begun to become popular, people's amateur cultural needs have increased significantly, and the dusty old things have aroused people's growing interest.

For the general public, under the circumstances that traditional historical books are difficult to understand in ancient times and modern history books of "academic mode" are difficult to "read," popularized history has almost become their only choice for "exploring the past."

Analyze the reasons and find the following: (1) the background material is discussed through the concept of "popular history fever" and many candidate sentences related to the question are not answer sentences, which need deep semantic understanding and reasoning technology. (2) It is found that there is a big semantic gap between "原因" in the question and the words such as "desire," "demand," "interest," and "choice" in the answer sentence. It is difficult for us to make semantic matching with existing tools such as HowNet, Word2Vector, and CFN.

The accuracy of extracting paragraph topic sentence and author's opinion sentence.

Annotate the paragraph topic sentences and author's opinion sentences on the Beijing 12 years College Entrance Examination. There are 19 materials, 89 paragraph topic sentences, and 26 author's opinion sentences. The experimental results are shown in Table 5.

Through the analysis of College Entrance Examination papers, it is found that, compared with the general news articles, it is more difficult to extract the topic sentence of the paragraph. As shown in Example 3, the topic sentence of the paragraph is "*Singing Kunqu Opera is something in the hall*" which is a concise summary of the paragraph. However, the similarity between topic sentences and other sentences is small, so it needs deeper semantic reasoning technology. The difficulty of extracting the author's opinion sentences is that some articles do not have a clear author's opinion. As shown in Example 3, the full text consists of four paragraphs. The first paragraph introduces "*Kunqu Opera*," and the next three paragraphs illustrate the strengths and limitations of "*Kunqu Opera*" from different perspectives, but there is no obvious general view and attitude.

Table 5: Experimental results of the paragraph topic sentence and author's opinion sentence.

Method	P (%)
Paragraph topic sentence recognition	80.62
Author's opinion sentence recognition	75.00

Example 3. 2009 Beijing College Entrance Examination

演唱昆曲是厅堂里的事情。地上铺了一方红地毯, 就算是剧中的境界, 唱的时候, 笛子是主要的乐器, 声音当然不会怎么响, 但是在一个厅堂里, 也就各处听得见了。搬上旧式的戏台去, 即使在一个并不宽广的戏院子里, 就不及平剧那样容易叫全体观众听清。如果搬上新式的舞台去, 那简直没有法子听, 大概坐在第五六排的人就只看见演员拂袖按鬓了。

Singing Kunqu Opera is something in the hall. There is a red carpet on the ground, even if it is the realm in the play; when singing, the flute is the main instrument, of course, the sound is not very loud, but in a hall, it can be heard everywhere. Moving on to an old-style theater, even in a theater that is not as wide as a theater, it is not as easy for the entire audience to hear. If you go to a new style stage, there is simply no way to listen. Perhaps the people sitting in the fifth and sixth rows will only see the actor's sleeves and temples.

4. Conclusion

After showing Gaokao's difficulty and its difference from the existing research problems, we propose a new framework for reading comprehension QA in Gaokao. The method first uses word similarity matching, frame matching, and discourse topic to construct the affinity matrix, which includes not only the relationship between the question and candidate sentences, but also the relationship between candidate sentences and then uses a graph-based algorithm to calculate the score of each sentence. Finally, the top 6 sentences are chosen as the answer sentences. At present, the deep reasoning ability of our method is not strong enough. In addition, the method in this article is extractive and cannot automatically generate some answers, so the score rate of the system is not high. In the next step, we will conduct a deep semantic understanding and reasoning on the background article and study a more efficient method. At the same time, we will further collect the relevant corpus to expand the scale of data and improve the answering effect of the system.

Data Availability

The data used to support the findings of this work are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Key R&D Projects (2018YFB1005103), the National Natural Science Foundation of China (61772324), and the 1331 Engineering Project of Shanxi Province of China.

References

[1] S. Guo, X. Zeng, S. He, K. Liu, and J. Zhao, "Which is the effective way for Gaokao: information retrieval or neural networks?" in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics EACL*, pp. 111–120, Valencia, Spain, April 2017.

[2] C. Gong, W. Zhu, Z. Wang, J. Chen, and Y. Qu, "Taking up the Gaokao challenge: an information retrieval approach," in *Proceedings of the 2016 International Joint Conference on Artificial Intelligence IJCAI*, pp. 2479–2485, New York, NY, USA, July 2016.

[3] A. Fujita, A. Kameda, K. Ai, and Y. Miyao, "Overview of Todai robot project and evaluation framework of its Nlp-based problem solving," in *Proceedings of the International Conference on Learning Representations ICLR*, pp. 2590–2597, Banff, Canada, April 2014.

[4] M. Feng, B. Xiang, M. R. Glass, L. Wang, and B. Zhou, "Applying deep learning to answer selection: a study and an open task," in *Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding ASRU*, pp. 813–820, Scottsdale, AZ, USA, December 2015.

[5] J. Chen, Qi Zhang, P. Liu, X. Qiu, and X. Huang, "Implicit discourse relation detection via a deep architecture with gated relevance network," in *Proceedings of the ACL*, pp. 1726–1735, Berlin, Germany, August 2016.

[6] L. Qin, Z. Zhang, H. Zhao, Z. Hu, and E. P. Xing, "Adversarial connective-exploiting networks for implicit discourse relation classification," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, pp. 1006–1017, Vancouver, Canada, July 2017.

[7] J. Cai, S. He, Z. Li, and H. Zhao, "A full end-to-end semantic role labeler, syntacticagnostic or syntactic-aware?" in *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, Santa Fe, NM, USA, August 2018.

[8] P. Rajpurkar, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the EMNLP 2016*, pp. 2383–2392, Association for Computational Linguistics, Austin, TX, USA, November 2016.

[9] W. He, "DuReader: a Chinese machine reading comprehension dataset from real-world applications," in *Proceedings of the MRQA 2018*, pp. 37–46, Melbourne, Australia, July 2018.

[10] Y. Cui, "A span-extraction dataset for Chinese machine reading comprehension," in *Proceedings of the EMNLP 2019 and 9th ICNLP*, pp. 5883–5889, Association for Computational Linguistics, HongKong, China, November 2019.

[11] D. Xiong, "A similarity calculation method of community question and answer based on LDA," *Journal of Chinese Information Processing*, vol. 26, no. 5, pp. 40–46, 2012.

[12] Z. Ye, "Research on open domain question answering system," in *Proceedings of the NLPCC 2015*, pp. 527–540, Springer, Nanchang, China, October 2015.

[13] L. T. Le, C. Shah, and E. Choi, "Assessing the quality of answers autonomously in community question–answering," *International Journal on Digital Libraries*, vol. 20, no. 4, pp. 351–367, 2019.

[14] C. Li, "Syntactic analysis and deep neural network in answer extraction of Chinese question answering system," *Journal of Chinese Mini-Micro Computer Systems*, vol. 38, no. 6, pp. 1341–1346, 2017.

[15] Q. Liu, "Semantic similarity of vocabulary based on HowNet," *International Journal of Computational Linguistics & Chinese Language Processing*, vol. 7, no. 2, pp. 59–76, 2002.

[16] W. T. Yih, "Question answering using enhanced lexical semantic models," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 1744–1753, Association for Computational Linguistics, Sofia, Bulgaria, August 2013.

[17] Y. Zhou, "A method of sentence semantic similarity based on synonym forest and its application in question answering system," *Computer Applications and Software*, vol. 36, no. 8, pp. 65–68, 2019.

[18] S. Guo, "Sentence semantic relevance for college entrance examination reading comprehension," *Journal of Tsinghua University (Science and Technology)*, vol. 57, no. 6, pp. 575–579, 2017.

[19] Y. Guan, "A study on the selection of text titles for Chinese reading comprehension in college entrance examination," *Journal of Chinese Information Processing*, vol. 32, no. 6, pp. 28–35, 2018.

[20] G. Li, "The extraction of answer sentences from Chinese reading comprehension of college entrance examination based on frame semantics," *Journal of Chinese Information Processing*, vol. 30, no. 6, pp. 164–172, 2016.

[21] L. Page, "The PageRank citation ranking: bringing order to the web," Technical Report, Stanford InfoLab, Stanford University, Stanford, CA, USA, 1999.

[22] C. Fan, *Research on PageRank Algorithm in Web Structure Mining*, Soochow University, Suzhou, China, 2nd edition, 2009.

[23] J. Liu, "Keyword extraction based on language network," in *Proceedings of the 3rd National Conference on Information Retrieval and Content Security NCIRCS 2007*, pp. 711–715, Soochow University, Suzhou, China, November 2007.

[24] X. Wan, "An exploration of document impact on graph-based multi-document summarization," in *Proceedings of the EMNLP 2008*, Association for Computational Linguistics, Honolulu, HI, USA, October 2008.

[25] M. Liu, *Sentence Similarity Calculation Based on Word Vector and Its Application in Case-Based Machine Translation*, Beijing Institute of Technology, Beijing, China, 2nd edition, 2015.

[26] R. Li, *Research on the Semantic Structure Analysis Technology of Chinese Sentence Frame*, Shanxi University, Taiyuan, China, 2nd edition, 2012.

[27] C. F. Baker, "The berkeley framenet project," in *Proceedings of the 17th ICCL*, pp. 86–90, Association for Computational Linguistics, Chicago, IL, USA, May 1998.

[28] HIT IR-Lab Tongyici Cilin (Extended), http://www.ir-lab.org/.

[29] W. Che, "Ltp: a Chinese language technology platform," in *Proceedings of the International Conference on Coling*, pp. 13–16, Association for Computational Linguistics, Beijing, China, August 2010.

[30] J. Devlin, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, Association for Computational Linguistics, Minneapolis, MN, USA, June 2019.