Answering Multiple-Choice Questions in Geographical Gaokao with a Concept Graph

Jiwei Ding, Yuan Wang, Wei Hu, Linfeng Shi, and Yuzhong Qu(✉)

National Key Laboratory for Novel Software Technology,
Nanjing University, Nanjing, China
jwdingnju@outlook.com, ywangnju@outlook.com, lfshinju@outlook.com,
{whu,yzqu}@nju.edu.cn

Abstract. Answering questions in Gaokao (the national college entrance examination in China) brings a great challenge for recent AI systems, where the difficulty of questions and the lack of formal knowledge are two main obstacles, among others. In this paper, we focus on answering multiple-choice questions in geographical Gaokao. Specifically, a concept graph is automatically constructed from textbook tables and Chinese wiki encyclopedia, to capture core concepts and relations in geography. Based on this concept graph, a graph search based question answering approach is designed to find explainable inference paths between questions and options. We developed an online system called CGQA and conducted experiments on two real datasets created from the last ten year geographical Gaokao. Our experimental results demonstrated that CGQA can generate accurate judgments and provide explainable solving procedures. Additionally, CGQA showed promising improvement by combining with existing approaches.

Keywords: Concept graph · Geographical Gaokao
Question answering · CGQA

1 Introduction

With the great development of Artificial Intelligence (AI) technologies, having machines to pass human examination is becoming a hot AI challenge. Similar to the Aristo project [5] and the Japanese Todai project [6], China has launched a National High-tech R&D Program [4] to promote AI systems to pass the national college entrance examination (commonly known as *Gaokao*). The goal of the project is to make AI systems not only answer complicated questions but also provide explainable solving procedures.

In recent years, a multitude of question answering (QA) approaches have been proposed for various application scenarios, such as reading comprehension [15], intelligent assistant [2], and community QA [12]. However, these approaches are not suitable for QA in Gaokao, due to the following three main difficulties:

A. Gangemi et al. (Eds.): ESWC 2018, LNCS 10843, pp. 161–176, 2018.
https://doi.org/10.1007/978-3-319-93417-4_11

- The difficulty in question understanding. Most multiple-choice questions in Gaokao use sentences as their options (answer choices) in order to test a student's ability in different topics. Figure 1 shows an example of multiple-choice questions in geographical Gaokao. In this example, option A is related to climate and vegetation in physical geography, while option B is related to agriculture in human geography. Also, there are multiple sentence structures in options, e.g., the first part of option C is factoid, while the second part contains a comparison. Traditional semantic parsing approaches [1,19] may have difficulty in processing such sentences, due to the lack of question patterns and specific knowledge bases.
- The lack of formal knowledge. Concepts play an important role for students to learn knowledge in high school, and form testing points in Gaokao questions. In the example question, "temperate steppe zone", "crop farming", etc., are all geographical concepts in textbooks. These concepts involve different topics and form complex hierarchies. Currently, geographical databases such as GeoNames[1] and Clinga [8] cover a large number of geographical entities. However, they contain little knowledge for geographical concepts and their relations.
- The complexity of inference. Inference is a common task in answering scientific questions, but the inference procedures in Gaokao are more complex. For example, option A needs student to judge whether London is "warm and rainy all year round". It is possible to solve this question by adapting related rules if we have London's temperature and precipitation data. However, students solve this question in another way, because "warm and rainy all year round" is a characteristic for the concept "temperate oceanic climate", which is London's climate type. For the first solution, AI systems may have difficulty in mapping natural language phrases to predicates and constants in rules, and need a large amount of geographical data. For the second solution, we need to collect related natural language descriptions for geographical concepts.

伦敦
A. 终年温和多雨，属于温带草原带　B. 农业以种植业为主，两年三熟
C. 受暖流影响，冬季气温高于同纬度地区　D. 冬至日有极夜现象

London (*translated into English*)
A. is warm and rainy all year round, belongs to the temperate steppe zone
B. agriculture is predominated by crop farming, three harvests per two years
C. is influenced by warm currents, winter temperature is higher than other areas at the same latitude
D. winter solstice has the polar night phenomenon

Fig. 1. A multiple-choice question in geographical Gaokao. To help readers understanding, the examples in this paper are translated into English.

In this paper, we focus on answering multiple-choice questions in geographical Gaokao. To overcome the above difficulties, we construct a concept graph

[1] http://www.geonames.org/ontology.

automatically, which not only captures the relations between geographical concepts, but also contains concept descriptions in natural language extracted from textbook tables and Chinese wiki encyclopedia. Furthermore, we propose a graph search based QA approach, which finds explainable inference paths between questions and options. We developed an online system called CGQA and conducted experiments on two real datasets created from the last ten year geographical Gaokao. Our experiments showed that CGQA generated accurate judgments and provided explainable solving procedures. Additionally, CGQA showed promising improvement by combining with existing approaches.

2 Related Work

Information retrieval techniques are wildly used in QA systems for short factoid questions. By searching sentences in a large web corpus, Clark et al. [5] scored 60.6% for multiple-choice questions in the 4th Grade Science Test. Cheng et al. [4] studied answering multiple-choice questions in historical Gaokao by retrieving and filtering pages in Chinese Wikipedia. However, the facts appeared in geographical Gaokao questions are not explicitly stated in text or databases, thus inference is required for this problem.

Over the years, several studies have been conducted to solve questions with inference. As an early study, OntoNova [3] represents chemistry knowledge in F-Logic to answer formal queries via rule reasoning. Similar approaches were used for physics and biology questions. These approaches cannot be directly applied to geographical Gaokao, since it is difficult to map natural language phrases to predicates and constants in rules. Khashabi et al. [9] employed textual entailment and integer liner programming to answer elementary science questions with tables. However, this approach relies on many manually-constructed constraints, and cannot handle questions with sentences as options.

Recently, deep neural networks achieve promising performance in some QA tasks. Sukhbaatar et al. [17] exploited an end-to-end memory networks, which outperformed other methods like LSTM on the synthetic QA tasks from Facebook. Guo et al. [7] proposed a permanent-provisional memory network to answer multiple-choice history questions in Gaokao, which gained the best score (45.70%) compared to other memory-capable neural network models. However, these approaches cannot provide explainable solving procedures, and sometimes suffer from the lack of training data.

In addition to a few open-domain knowledge bases like DBpedia [10] and XLore [18], GeoNames is perhaps the most well-known database for geographical information, which contains data like place names, latitude and longitude coordinates. Clinga [8] is a Chinese linked geographical dataset, which contains a large number of entities extracted from Chinese wiki encyclopedia, and links them to other knowledge bases like DBpedia. All datasets mentioned above cover a large number of geographical entities such as cities and mountains, but contain little knowledge of geographical concepts, which wildly appear in geographical Gaokao questions.

With the rapid development of KBQA methods, users are able to query knowledge bases with natural language questions. Zou et al. [19] exploited semantic query graphs for question analysis. Abujabal et al. [1] mapped questions to automatically generated SPARQL templates. However, these approaches may have difficulty in processing Gaokao questions, due to the lack of question patterns and specific knowledge bases.

3 Concept Graph Construction

We first give the definition for the concept graph as follows:

Definition 1 (Concept Graph). *We define a concept graph as a 5-tuple* (S, C, D, R, T)*, where* S, C, D *and* R *denote the sets of concept schemes (topics), concepts, descriptions and relations, respectively.* $T \subseteq (C \times R \times S) \cup (C \times R \times C) \cup (C \times R \times D)$ *denotes the set of triples.*

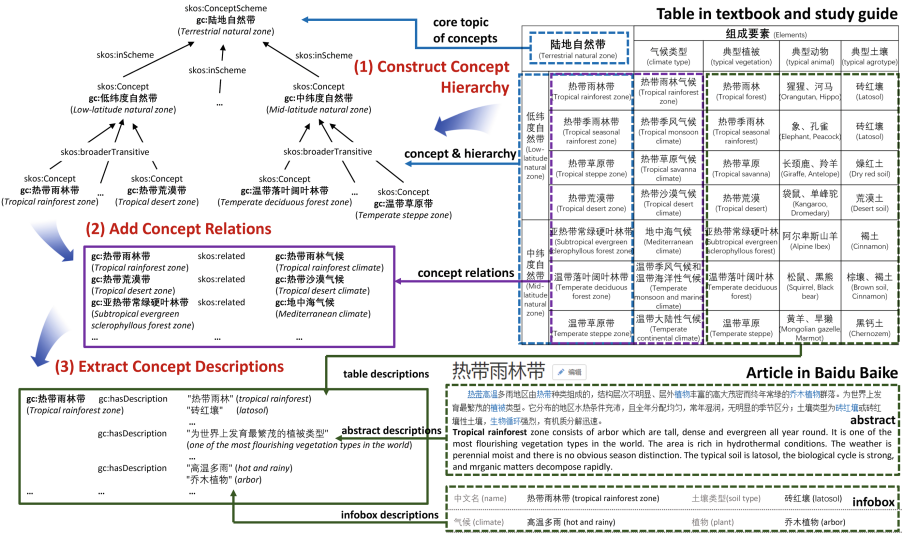

Fig. 2. General steps to construct concept graph for geographical Gaokao

We construct our geographical Gaokao concept graph according to Simple Knowledge Organization System (SKOS). Figure 2 depicts the general steps to construct the concept graph, including: (1) Construct Concept Hierarchy, (2) Add Concept Relations, and (3) Extract Concept Descriptions. The methodology that we use is largely automatic, with little manual amendment to ensure quality.

3.1 Construct Concept Hierarchy

We choose geography textbooks and study guides[2] as the main data sources due to their high quality and targetedness. These materials usually organize different concepts in the form of tables in addition to plain text descriptions, to help students to explicitly notice their commodities and differences. For a table in textbook, geographical concepts are often placed in the first column, with the concepts' topic name (e.g., Terrestrial natural zone) in the column header. For each concept, a unique ID is generated as URI (e.g., *gc:Tropical_rainforest_zone*) and *skos:Concept* is defined as its type. Similarly, a URI is generated for the concepts' topic name and is declared with type *skos:ConceptScheme*. *skos:inScheme* relation is used to connect concepts with their topics. The broader-narrower relations between concepts can also be extracted from the table cell hierarchy, and are represented using *skos:broaderTransitive* and *skos:narrowerTransitive* in concept graph. If two concepts in the same *skos:ConceptScheme* do not have the broader-narrower relation, a *gc:disjointWith* relation is added between them automatically. A few tables have empty column headers, or use common words such as "type" as their column headers. This issue is manually fixed by giving an appropriate topic name.

3.2 Add Concept Relations

Geographical concepts in different topics often relate to each other (e.g., *volcanic landform* is related to *magmatic rock*, *tropical rainforest zone* is related to *tropical rainforest climate*). We collect this kind of relations through the co-occurrence of concepts in the following sources and represent them using relation *skos:related*.

- Textbook. Concepts appearing in the same row of a table are considered as related. Also, if two concepts co-occur frequently in text, we consider them as related. In our construction process, this frequency threshold is set to 2 based on our experimental experience.
- Baidu Baike. Baidu Baike is the largest collaboratively-built Chinese wiki encyclopedia. The first few paragraphs are an overview of the geographical concept, called *abstract*. We consider two concepts as related if both of them appear in each other's Baidu Baike abstract.

It is worth noting that *skos:related* is only used between concepts in different concept schemes, as concepts in the same concept scheme only have *skos:broaderTransitive*, *skos:narrowerTransitive* or *gc:disjointWith* relation.

3.3 Extract Concept Descriptions

Many concepts have specific natural language phrases describing their characteristics of different aspects. For example, "hot and rainy" describes the climatic characteristic for *gc:Tropical_rainforest_zone*. We call these phrases *Descriptions*, and store them in plain text as the values of *gc:hasDescription*. Descriptions are automatically extracted from the following sources:

[2] We used electronic editions at http://kb.cs.tsinghua.edu.cn/res/index.

– Textbook tables. Table cells in the same row of a concept can be regarded as its descriptions if they do not contain any other concepts.
– Baidu Baike infobox and abstract. In a concept's Baidu Baike article, the key-value pairs in the infobox present some structural facets of the concept. We consider each value as a description, if it does not contain any other concepts. Especially, values of "alias" and "also known as" are described as the value of *skos:altLabel*. Also, each sentence in the abstract is considered as a description if it has proper length (less than 30 Chinese characters in our current setting) and does not contain any other concepts. We will consider the method in [16] to extract descriptions from tables in Baidu Baike in the future.

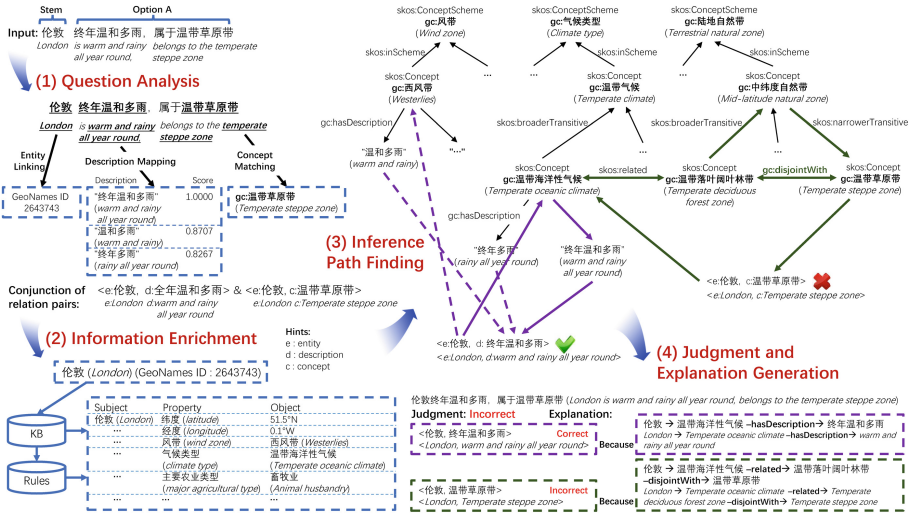

Fig. 3. Multiple-choice question answering with concept graph

4 Multiple-Choice Question Answering

A multiple-choice question in geographical Gaokao contains a string qs called *question stem* and four strings $O = \{o_1, o_2, \ldots, o_4\}$ called *options*. Some questions contain background text qb or diagrams qd as extra materials.

The framework of our QA approach is shown in Fig. 3. Given a multiple-choice question, we convert it to four statements by combining each option o_i with the question stem qs. Our approach takes each statement as input, and gives judgment and explanation with four stages: (1) Question Analysis, (2) Information Enrichment, (3) Inference Path Finding, and (4) Judgment and Explanation Generation. The statement with the highest score among the four is chosen as the answer.

4.1 Question Analysis

The question analysis stage in our approach contains three steps: (1) Entity Linking and Concept Matching, (2) Description Mapping, and (3) Relation Pairs Generation.

Entity Linking and Concept Matching. In geographical Gaokao, entities and concepts usually appear with their canonical names to ensure the rigor of the exam. This makes the linking process easy. We use a sliding window based approach to recognize concept and entity mentions, and link them to the concept graph and GeoNames, respectively. For entities with the same name in GeoNames, we select the one with the highest administrative division level and the latest update date. Furthermore, a few patterns are manually built to recognize anonymous entities such as "place A" and "area One", which are often used to refer to entities in figures or background materials. Our experiment over 200 multiple-choice questions shows that, by using the above method, the precision and recall reached 92% and 90%, respectively.

Description Mapping. Description mapping is to recognize the descriptive text in question stems and options, and map them to the descriptions (D) in the concept graph, which is quite similar with the semantic matching task [11]. Each question and option is split into several clauses according to conjunction and punctuation, and each clause is supposed to contain at most one continuous descriptive text, which might map to several descriptions in the concept graph. For example, the option A in Fig. 3 is split into two clauses, and the first one is mapped to description "warm and rainy all year round" and "warm and rainy". Since there are thousands of descriptions in the concept graph, we firstly tried a Lucene-based approach to ensure efficiency. However, this approach failed to achieve satisfactory mapping precision since Lucene only considers lexical similarity and TF-IDF value. In order to achieve higher mapping precision, a two-step mapping approach is developed.

Firstly, we use a comprehensive measurement which considers both Levenshtein distance and word embedding similarity to find candidate descriptions for each clause. The Levenshtein distance measures the lexical similarity, while the word embedding similarity measurement handles the problem of semantic heterogeneity. Each n words in the clause (n decreases from clause length to 1) is seen as a query q, and the mapping score between query q and description d is calculated as follows:

$$MScore(q,d) = \alpha\Big(1 - \frac{LevenshteinDistance(q,d)}{\max(Length(q), Length(d))}\Big) + (1-\alpha)\cos(v(q), v(d)), \tag{1}$$

where $v(q)$ and $v(d)$ stand for the embedded word vectors for query and description, respectively. The training process of the word embedding model is introduced in Sect. 5.4. The longest query that has mapping score larger than θ with any description is seen as the descriptive text for the clause, and all descriptions having mapping scores larger than θ are seen as candidate descriptions.

Secondly, we find that adjectives usually play an important role when measuring semantic similarity. For example, "warm and rainy all year round" is similar to "cold and rainy all year round" in lexical similarity, but they refer to totally different climates in geography. In our approach, we exploit adjective relations in WordNet [14] to filter candidate descriptions. Suppose that candidate description d contains several adjectives $d_1, \ldots, d_i, \ldots, d_N$, and the query q contains adjectives $q_1, \ldots, q_j, \ldots, q_M$. If any pair of d_i and q_j have the antonym relation in WordNet, description d is removed from the candidates. After filtering, the top k candidate descriptions are selected as the output of our description mapping approach.

Relation Pairs Generation. In this step, entities, concepts and descriptions are reorganized as relation pairs according to their relationships in the dependency tree. Particularly, some relation pairs are ignored since both of their components are in the question stem. Then, the whole input statement is transformed into the conjunction of relation pairs. The disjunction of relation pairs is not considered in our work, since it never appears in real-life Gaokao questions. An example of the conjunction of relation pairs is as follows: $\langle e : London, d : warm\_and\_rainy\_all\_year\_round \rangle$ & $\langle e : London, c : Temperate\_ste\text{-}ppe\_zone \rangle$, where e: stands for entity, c: stands for concept, and d: stands for description.

4.2 Information Enrichment

Inferring concepts associated with a given geographical entity is an important ability for students, and is frequently tested in Gaokao (e.g., inferring the climatic zone of London). We implement two different methods to relate geographical entities to concepts and descriptions in the concept graph.

- **Structured Knowledge Acquisition**. Knowledge bases such as GeoNames (see footnote 1), Wikidata[3], Clinga [8] and Koppen[4] provide plenty of geographical information for entities, such as latitude, climate type and precipitation. Entities can be related to concepts directly by querying the knowledge bases, or indirectly by applying geographical rules. For example, "The latitude of westerlies is between 35 and 65°" is a simple geographical rule that links entity and wind belt in the concept graph using latitude information, which can be fetched in GeoNames and Wikidata for most of the entities.
- **Related Text Matching**. Some geographical concepts and descriptions occur in entities' related text, such as background materials in a test paper, or articles in Chinese wiki encyclopedia. We use the mapping approach described in the question analysis section to extract related concepts and descriptions from these materials. Also, the concepts having "examples" or "typical areas" as their descriptions may contain the entity names directly.

[3] https://www.wikidata.org/.
[4] http://hanschen.org/koppen/.

4.3 Inference Path Finding

Our approach infers *supporting paths* and *protesting paths* in the concept graph, which are defined as follows:

Definition 2 (Supporting Path). *A path is a supporting path if and only if no gc:disjointWith relation exists between any two vertices in this path. Formally, let G be a concept graph. Given a path P containing vertices $V(P) = \{v_1, v_2, \ldots, v_n\}$, P is a supporting path iff $\nexists v_i, v_j \in V(P)$, $v_i \xrightarrow{gc:disjointWith} v_j \in G$.*

Definition 3 (Protesting Path). *A path is a protesting path if and only if at least a gc:disjointWith relation exists between some of the vertices in this path. Formally, let G be a concept graph. Given a path P containing vertices $V(P) = \{v_1, v_2, \ldots, v_n\}$, P is a protesting path iff $\exists v_i, v_j \in V(P)$, $v_i \xrightarrow{gc:disjointWith} v_j \in G$.*

The following theorem is easy to be proved.

Theorem 1 *A path in the concept graph is either a supporting path or a protesting path.*

The procedure for making the judgment for a relation pair $\langle A, B \rangle$ is shown below:

1. Add nodes A and B to the concept graph, connect them with related concepts and descriptions.
2. Find a supporting path from A to B using Algorithm 1. If found, return **Correct**.
3. Otherwise, find a shortest path from A to B through the Breadth-First Search. If found, according to Theorem 1, this path must be a protesting path, return **Incorrect**.
4. If there is no path from A to B in the concept graph, return **Unknown**.

Algorithm 1. Supporting path finding

Input: Nodes in relation pair $\langle A, B \rangle$ and concept graph G
Output: A path from A to B, or null for no path found

```
function FindSupportPath(A, B, G)
    if A = B then
        return a path only containing B;
    G' ← G − {v | v --gc:disjointWith--> A, v ∈ G};
    for all A's neighbour v_i in G' do
        P ← FindSupportPath(v_i, B, G' − {A});
        if P ≠ null then
            return append A to the head of path P;
    return null;
```

4.4 Judgment and Explanation Generation

Our approach generates the judgment for each statement by combining the judgment for each relation pair through the following method:

- If any relation pair in the statement is incorrect, the whole statement is incorrect, scores -1;
- If there are m correct relation pairs and n unknown relation pairs in the statement, the whole statement is partial correct, scores $\frac{m}{n+m}$;
- If all the relation pairs of the statement are unknown, the whole statement is not judged, scores 0;

Additionally, our approach combines the inference path for each relation pair to generate the explanation for the whole statement. For a correct relation pair, we display the supporting path found by Algorithm 1; while for an incorrect relation pair, we display the shortest protesting path. Relation pairs with no judgment are ignored to explain.

Table 1. Statistics of the geographical Gaokao concept graph

	Names	Count
Nodes	*skos:Concept Scheme*	57
	skos:Concept	588
	Plain literals	4,312
Edges	*skos:inScheme*	588
	skos:broader Transitive	373
	skos:narrower Transitive	373
	skos:related	142
	gc:disjointWith	12,602
	skos:prefLabel	645
	skos:altLabel	213
	gc:hasDescription	3,453

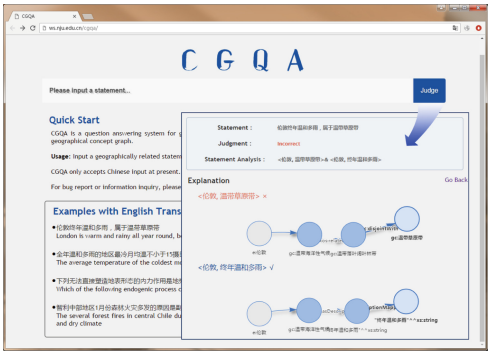

Fig. 4. Screenshot of CGQA

5 Experiments

5.1 Geographical Gaokao Datasets

We collected multiple-choice questions with expected answers from real-life geographical Gaokao all over the states in recent ten years (2008–2017). After removing duplicate questions, 4,305 multiple-choice questions were collected in total: 1,756 from 128 geography tests in Gaokao, and 2,549 from 116 geography tests in mock Gaokao. The average length for question stems and options are 19.65 and 8.44, respectively. About 87% of questions contain diagrams, which are currently

difficult to be processed. To evaluate our approach, we constructed the following two datasets [5]:

- **Beijing Geographical Gaokao (BGG):** It contains all multiple-choice questions in Beijing Geography Gaokao from 2008 to 2017: six no-diagram questions and 104 questions with 60 different diagrams. All of these diagrams were annotated by geography majors in triples in advance. These labels only contain basic information for anonymous entities, such as longitude and latitude, which can only be fetched from diagrams.
- **No-Diagram Questions (NDQ):** It contains all no-diagram multiple-choice questions in Gaokao and mock Gaokao all over the states from 2008 to 2017. There are 545 questions in total.

5.2 Geographical Gaokao Concept Graph

Table 1 shows the numbers of nodes and edges of our concept graph [6]. In total, it contains 19,034 triples describing 588 concepts in 57 concept schemes.

To evaluate the accuracy of concept relations and descriptions, we manually judged the correctness of all *skos:related* relations and 500 randomly picked descriptions. The precisions are 93.66% and 90.60%, respectively. We observed that the precision of descriptions extracted from Baidu Baike infoboxes is not good, because some contain short common words such as "global" and "geography" as values.

5.3 Demo

The demo of our approach, called CGQA, is currently available at http://ws.nju.edu.cn/cgqa/. As shown in Fig. 4, the system will give out judgment and solving procedure after user input a geographically related statement.

5.4 Comparative Approaches

We compared our approach with three existing approaches. Each approach is required to provide a score for each option, and the options with the highest score are chosen as the answer. For our approach, the parameter α in Eq. (1) is set to 0.2, the mapping score threshold θ is set to 0.80, and k is set to 3 in description mapping.

IR-Based Approach. The information retrieval (IR) based approach [4] is to find out the confidence that the question stem qs along with an answer option o_i is explicitly stated in a corpus. For this purpose, a text corpus containing 14.8 million sentences was automatically built using geographical textbooks, study guides and Baidu Baike articles used in Clinga. We used $qs + o_i$ as the input for Lucene, and returns the Lucene's score for top retrieved sentence having at least one non-stopword overlap with qs and o_i. This is repeated several times to score all options.

[5] Both datasets are available at http://ws.nju.edu.cn/cgqa/datasets.zip.
[6] The concept graph is available at http://ws.nju.edu.cn/cgqa/cg.zip.

WE-Based Approach. The word embedding (WE) based approach [5] computes the semantic relevance between each answer option o_i and the question stem qs by exploiting word similarity. In our approach, a word embedding (150 dimensions) was learned using Skip-gram [13] over the corpus built for the IR-based approach. We defined the semantic relevance of qs and o_i as the cosine similarity between their composite vectors, which were computed by summing the vector for each word in qs and o_i, respectively.

NN-Based Approach. As neural network (NN) is widely used in QA, we designed an end-to-end NN-based approach according to [7,17]. All questions not included in the test set were treated as training data (3,650 questions in total). The design of our network is shown in Fig. 5, which includes:

- The input module maps the question stem, options, and the top 10 related sentences retrieved by Lucene into vectors using a pre-trained word2vec model (the same as the WE-based approach);
- The encoder module embeds the vectors from each input to a new vector space by a single Bi-GRU layer;
- The attention module combines the top 10 related vectors using weighted sum, and adds the combined vector with the question stem vector as the final question vector;
- The output module computes the match between the final question vector and option vectors by taking the inner product followed by a softmax.

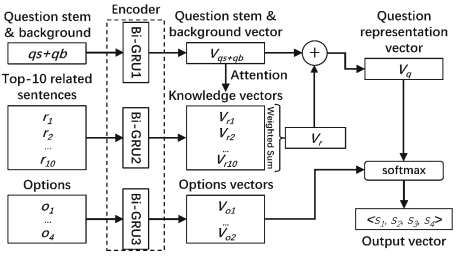

Fig. 5. Design of NN-based approach

Table 2. Results for accuracy of judgments and explanations

	BGG	NDQ
Number of statements	440	2,180
Answered percentage	11.36%	11.24%
Judgment precision	84.00%	90.61%
Explanation correctness	81.20%	88.52%
Average path length	3.76	3.05

5.5 Procedures and Metrics

We firstly evaluated the accuracy of judgments and explanations of our approach. For each multiple-choice question, four statements were generated by combining stem text with each option. The following metrics were used in this step:

- Answered percentage: percentage of judged statements in all statements;

- Judgment precision: percentage of correctly-judged statements in all judged statements;
- Explanation correctness: percentage of correct explanations in all the explanations, to avoid the statement being answered correct by accident.

To measure the explanation correctness, five undergraduate students of geography specialty were invited to score the explanations provided by our approach. An explanation is labeled correct only if it provides enough evidence to make the accurate judgment. Explanations for wrong judgments were labeled incorrect automatically.

Secondly, we compared the scores of our approach with several comparative approaches listed above. In addition, we combined our approach with each comparative approach using priority strategy to observe whether there are scores improvements. For scoring function, we followed the work in [5]. If the approach provides N answers including the correct one, it scores $1/N$. If no answer is produced, it scores $1/K$ for question with K options, equivalent to random guessing.

5.6 Results

Table 2 shows the accuracy of judgments and explanations for our approach. Our approach judged a part of statements (about 11%) with high precision (84.00% and 90.61% for BGG and NDQ, respectively). The reason of unsatisfactory answered percentage is that there are more than 60% statements that do not contain any concept or description in both two datasets. The average length of the inference paths is larger than 3, which indicates that these questions are quite difficult. Additionally, the mean value for explanation correctness reaches 81.20% and 88.52%, respectively, and the standard deviation between different assessors is approximately 1%. Repeated Measures ANOVA indicates that there is no divergence between different assessors ($p > 0.95$). Most assessors reported that the inference path proposed by our approach is reasonable and easy to be understood during the experiments.

Table 3 shows the test performance for each single approach. Our approach outperformed the IR-based and WE-based approaches, and achieved comparable results with the NN-based approach. When we took a look at the output for each approach, we found that our approach only answered a few part of the questions (22.73% of BGG and 16.88% of NDQ). The score of our approach on these answered questions is 68.00% and 83.33%, respectively. In contrast, all the comparative approaches answered nearly 100% of the questions, with scores lower than 38%.

In Table 4, comparative approaches gained an increment of approximately 6%–14% in test scores by combining with our approach. This indicates that our approach and existing approaches are complementary. CGQA + NN achieved the best performance on both datasets (43.41% and 45.50%, respectively).

Additionally, we allowed each comparative approaches to answer a part of questions according to the descending order of confidence value, where confidence

Table 3. Scores for single approaches

	CGQA	IR	WE	NN
BGG	**34.77%**	20.91%	22.73%	32.50%
NDQ	34.85%	28.99%	29.72%	**37.25%**

Table 4. Scores for combined approaches

	CGQA + IR	CGQA + WE	CGQA + NN
BGG	30.76%	36.36%	43.41%
NDQ	35.23%	39.45%	45.50%

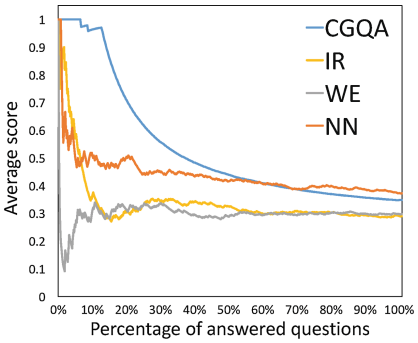

Fig. 6. Scores for single approaches

value is defined as the difference between the highest score and the second-highest score among the options. Figure 6 shows the average score on answered questions for these approaches on NDQ. It can be found that, the score of IR-based approach declined quickly when answered percentage becomes larger, which means only a few question-answer pairs can be directly found in corpus. Although the NN-based approach achieved highest overall precision, its average score on top 10% high confidence questions is lower than 50%, which indicates that the result might be unstable and inexplainable. Our CGQA approach achieved extremely high score (nearly 100%) on questions with high confidence value, and performed best when answering less than 60% questions.

6 Discussion

The experimental results allow us to make the following observations:

- CGQA achieved high scores on a part of the questions and provided explainable solving procedures. The average length of inference path is larger than 3, which proves the difficulty of these questions. However, the coverage of our approach is restricted to the scale and quality of the concept graph. Extracting knowledge from web tables and plain text in textbooks should be considered in the future.
- There are still some questions cannot be solved with our current CGQA architecture, such as questions with spatio-temporal reasoning and questions involving numerical calculation. A question classification stage will be needed to make better combination with upcoming QA approaches.
- CGQA can also be adapted to answer short answer questions by searching relevant concepts and descriptions in the concept graph. At the time of writing this paper, we are working on transforming inference paths in the concept graph to natural language question answers. A challenge that we are facing is analyzing background materials in short answer questions. There might

be some newly-defined concepts or introduced rules in the material, which may lead to dynamic adjustment of the concept graph and the path finding algorithm.
– At present, we only consider geographical Gaokao due to our project goal. In the future, we are going to apply CGQA to other subjects such as history. Most of the concepts in history can also be organized as tree hierarchies, such as "political system" and "school of thought", which indicates that our CGQA approach can still be feasible.

7 Conclusion

In this paper, we studied the problem of answering multiple-choice questions in geographical Gaokao. Our main contributions are summarized as follows:

– We constructed a concept graph of high quality from textbook tables and Chinese wiki encyclopedia, to capture core concepts and relations in geography. The largely-automated construction approach can be applied to other domains.
– We proposed a graph search based QA approach to find explainable inference paths between questions and answer choices.
– We developed an online system CGQA and conducted experiments on real datasets created from the last ten year geographical Gaokao. Our experiments showed that CGQA generated accurate judgments and provided explainable solving procedures. Also, CGQA gained promising improvement by combining with existing methods.

Acknowledgments. This work is funded by the National Natural Science Foundation of China (No. 61772264) and the National High-tech R&D Program of China (No. 2015AA015406). We thank all participants in the evaluation for their time and effort.

References

1. Abujabal, A., Yahya, M., Riedewald, M., Weikum, G.: Automated template generation for question answering over knowledge graphs. In: Proceedings of the 26th International Conference on World Wide Web, pp. 1191–1200. International World Wide Web Conferences Steering Committee (2017)
2. Agrawal, R., Golshan, B., Papalexakis, E.: Toward data-driven design of educational courses: a feasibility study. J. Educ. Data Min. **8**(1), 1–21 (2016)
3. Angele, J., Moench, E., Oppermann, H., Staab, S., Wenke, D.: Ontology-based query and answering in chemistry: OntoNova Project Halo. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 913–928. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-39718-2_58
4. Cheng, G., Zhu, W., Wang, Z., Chen, J., Qu, Y.: Taking up the Gaokao challenge: an information retrieval approach. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, pp. 2479–2485. AAAI (2016)
5. Clark, P., Etzioni, O., Khot, T., Sabharwal, A., Tafjord, O., Turney, P.D., Khashabi, D.: Combining retrieval, statistics, and inference to answer elementary science questions. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, pp. 2580–2586. AAAI (2016)

6. Fujita, A., Kameda, A., Kawazoe, A., Miyao, Y.: Overview of Todai robot project and evaluation framework of its NLP-based problem solving. In: Proceedings of the 9th International Conference on Language Resources and Evaluation, pp. 2590–2597 (2014)
7. Guo, S., Zeng, X., He, S., Liu, K., Zhao, J.: 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, pp. 111–120. ACL (2017)
8. Hu, W., Li, H., Sun, Z., Qian, X., Xue, L., Cao, E., Qu, Y.: Clinga: bringing Chinese physical and human geography in linked open data. In: Groth, P., Simperl, E., Gray, A., Sabou, M., Krötzsch, M., Lecue, F., Flöck, F., Gil, Y. (eds.) ISWC 2016. LNCS, vol. 9982, pp. 104–112. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46547-0_11
9. Khashabi, D., Khot, T., Sabharwal, A., Clark, P., Etzioni, O., Roth, D.: Question answering via integer programming over semi-structured knowledge. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, pp. 1145–1152. AAAI (2016)
10. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., Bizer, C.: DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. Semant. Web J. **6**(2), 167–195 (2015)
11. Li, H., Xu, J.: Semantic matching in search. Found. Trends Inf. Retr. **7**(5), 343–469 (2014)
12. Lu, H., Kong, M.: Community-based question answering via contextual ranking metric network learning. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence, pp. 4963–4964. AAAI (2017)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of the 1st International Conference on Learning Representations (2013)
14. Miller, G.A.: WordNet: an electronic lexical database. Commun. ACM **38**(11), 39–41 (1995)
15. Richardson, M., Burges, C.J.C., Renshaw, E.: MCTest: a challenge dataset for the open-domain machine comprehension of text. In: Proceedings of the 20th Conference on Empirical Methods in Natural Language Processing, pp. 193–203. ACL (2013)
16. Ritze, D., Lehmberg, O., Oulabi, Y., Bizer, C.: Profiling the potential of web tables for augmenting cross-domain knowledge bases. In: Proceedings of the 25th International World Wide Web Conference, pp. 251–261. ACM (2016)
17. Sukhbaatar, S., Szlam, A., Weston, J., Fergus, R.: End-to-end memory networks. In: Proceedings of the 29th International Conference on Neural Information Processing Systems, pp. 2440–2448. Curran Associates, Inc. (2015)
18. Wang, Z., Li, J., Wang, Z., Li, S., Li, M., Zhang, D., Shi, Y., Liu, Y., Zhang, P., Tang, J.: XLore: a large-scale English-Chinese bilingual knowledge graph. In: Proceedings of the 12th International Semantic Web Conference, pp. 121–124 (2013)
19. Zou, L., Huang, R., Wang, H., Yu, J.X., He, W., Zhao, D.: Natural language question answering over RDF: a graph data driven approach. In: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, pp. 313–324. ACM (2014)