Towards Answering Geography Questions in Gaokao: A Hybrid Approach*

Zhiwei Zhang, Lingling Zhang, Hao Zhang, Weizhuo He, Zequn Sun, Gong Cheng, Qizhi Liu, Xinyu Dai, and Yuzhong Qu

Nanjing University, 163 Xianlin Avenue, Nanjing 210023, Jiangsu, China
{zwzhang,llzhang,haozhang,heweizhuo,zequnsun}@smail.nju.edu.cn
{gcheng,lqz,daixinyu,yzqu}@nju.edu.cn

Abstract. Answering geography questions in a university's entrance exam (e.g., Gaokao in China) is a new AI challenge. In this paper, we analyze the difficulties in problem understanding and solving, suggesting the necessity of developing novel methods. We present a pipeline approach which mixes information retrieval techniques with knowledge engineering and exhibits an interpretable problem solving process. Our implementation integrates question parsing, semantic matching, and spreading activation over a knowledge graph to generate answers. We report its promising performance on a representative sample of 1,863 questions used in real exams. Our analysis of failures reveals a number of open problems to be addressed in the future.

Keywords: Information retrieval · Knowledge engineering · Natural language processing· Question answering

1 Introduction

Subsequent to the Japanese Todai project [4], China has launched an ambitious initiative to have AI pass the National Higher Education Entrance Examination, commonly known as Gaokao. To meet this grand challenge, advances in AI technologies can in the meantime benefit future-generation computer-aided education, e.g., enabling applications that can interact with learners and act as mentors or provide assistance to teachers .This long-term goal drives our research of an approach that is capable of not only correctly answering complex questions but also solving problems like human intelligence, i.e., to have an interpretable problem solving process.

Previous research focused on history exams in Gaokao [2, 13]. In this paper, we present the first study on geography exams. An example of geography questions is illustrated in Fig. 1. Geography questions appear more difficult than history questions for both students and machines, due to not only the understanding and exploitation of graphical presentations (e.g., maps and charts) but

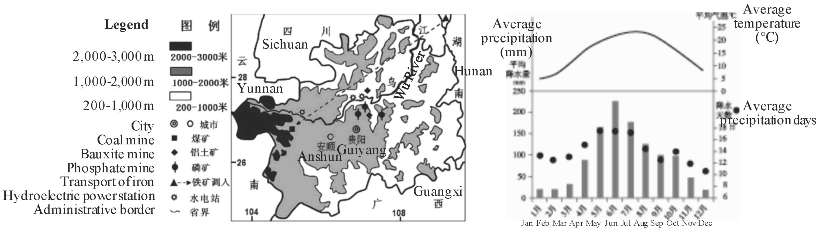
---

* This work was supported in part by the 863 Program under Grant 2015AA015406, in part by the NSFC under Grant 61572247 and 61772264, and in part by the Qing Lan and Six Talent Peaks projects.

also the complexity and diversity in problem solving. These distinguishing features have turned previous solutions to history questions and existing methods for question answering (QA) ineffective. Specifically, our investigation of more than one thousand real geography questions finds that: quantitative methods (e.g., math calculation) and qualitative methods (e.g., text retrieval) are almost equally important; most questions require mixing them together in a hybrid reasoning process.

Therefore, we propose a hybrid approach which has a pipeline design to enable interpretable problem solving. Our contributions in this first attempt are threefold.

- We collect, investigate, and make available 1,863 geography questions used in Gaokao or mock exams in recent years. In Section 2, we survey the difficulties in answering these questions and summarize the new technical challenges they bring to AI, which can hardly be met by existing methods.
- We present our approach which mixes information retrieval techniques with knowledge engineering. Section 3 gives an overview of its pipeline, and Section 4 describes its current implementation consisting of question parsing, knowledge graph construction, semantic matching, spreading activation and answer generation.
- We carry out extensive experiments on the collected questions, and create a gold standard for the output of each component of the proposed approach. In Section 5 we evaluate the approach and its components, show their effectiveness, and analyze their failures.


Legend
图 例
2,000-3,000m
2000-3000米
1,000-2,000m
1000-2000米
200-1,000m
200-1000米
City
城市
Coal mine
煤矿
Bauxite mine
铝土矿
Phosphate mine
磷矿
Transport of iron
铁矿调入
Hydroelectric power station
水电站
Administrative border
省界
四 川
Sichuan
云
28
Yunnan
江
湖
Hunan
乌
Wu River
南
安顺
Anshun
贵阳
Guiyang
26
南
广
Guangxi
西
104
108
Average precipitation (mm)
平均降水量
Average temperature (°C)
平均气温℃
Average precipitation days
降水天数
250
200
150
100
50
0
25
20
15
10
5
0
18
16
14
12
10
8
6
1月
2月
3月
4月
5月
6月
7月
8月
9月
10月
11月
12月
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec

贵州地形与主要矿产资源分布示意图
Terrain and major mineral resources in Guizhou

贵阳气候基本数据统计图
Basic climate data for Guiyang

Questions:

(1) 描述贵州省的地理位置特点。
Q1: Describe the features of the location of Guizhou Province.

(2) 从地形、气候角度说明自然条件对贵州经济发展的不利影响。
Q2: In terms of terrain and climate, explain the negative effects of natural circumstances on the economic development of Guizhou.

(3) 2014年1月，国家级开发区——贵安新区在贵阳和安顺之间正式成立。我国三大电信运营商在此设立云计算中心，中关村科技园、搜狐、百度、京东等企业落户此地，贵安大数据产业基地初具雏形。根据该材料概括贵安大数据产业基地赢得的有利发展条件，结合所学指出该产业基地为吸引企业入驻应进一步采取的措施。
Q3: In January 2014, the Gui-An state-level new area was established, located between Guiyang and Anshun. Over there, the three state-run telecom companies opened their cloud computing centers, and companies including Centek, Sohu, Baidu, and JD set up new branches, showing the initiation of the Gui-An base for the big data industry. With respect to this information, summarize the positive circumstances for developing the Gui-An base for the big data industry, and identify some measures for this industrial base to attract more companies.

Fig. 1: An example of geography questions in Gaokao based on a map and a chart.

2 Challenges and Related Work

Differences between Geography and Other Subjects. Research attention has been given to complex questions appearing in entrance exams at different levels of education including Gaokao. The OntoNova system [1] represents chemistry knowledge in F-Logic, to answer formal queries regarding complex chemical concepts and reactions via rule reasoning. Similar approaches are used to answer physics and biology questions [5]. Different from these natural or formal sciences which require a fully formal method with an emphasis on deep reasoning, history [2] questions are solved primarily by using information retrieval techniques supplemented by shallow reasoning.

Geography, broadly consisting of physical geography and human geography, is distinguished from the above-mentioned subjects in problem understanding and solving. As illustrated in Fig. 1, geography questions are unique because they always involve maps, charts, or other graphical presentations from which information is to be extracted as it is usually essential to problem solving. The understanding of the natural language expressing those questions (e.g., Q1–Q3 in Fig. 1) is distinctively challenged by the lack of a vocabulary of geographical concepts (e.g., *terrain* and *climate* in Q2), the extensive use of unbounded commonsense knowledge (e.g., *state-level new area* and *big data industry* in Q3), and many new question types that are specific to geography. To solve those questions which often exhibit a hybrid nature, the techniques we need include (i) all kinds of knowledge representations (e.g., ontological, causal, spatial, and temporal relations) and reasoning capabilities (e.g., logical reasoning, math calculation) which have only been partially considered in answering physics and chemistry questions, (ii) qualitative methods (e.g., text retrieval and summarization) which have only been preliminarily used in answering history questions, and more importantly, (iii) an effective mixture of these methods within a single problem solving process.

Challenges to Traditional Methods for QA. Geography questions in Gaokao are even further beyond the capabilities of traditional methods for QA. None of the above-mentioned challenges can be met by early information retrieval solutions [8] or recent semantic parsing techniques [9] and QA systems [3]. In fact, those methods primarily focus on retrieving answers to short factoid questions, thereby not fitting geography questions in Gaokao. As illustrated by Q3 in Fig. 1, geography questions in Gaokao can be too long to be processed by existing QA solutions, and are often non-factoid, expecting answers that are not a word or phrase but several generated sentences. Other end-to-end methods [11] suffer from lacking interpretability and hence diverge from the long-term goal of our research.

To conclude, not only novel techniques for QA are necessary for effective and interpretable solving of geography questions, but also an effort of domain-specific knowledge engineering seems unavoidable.

3 Overview of the Approach

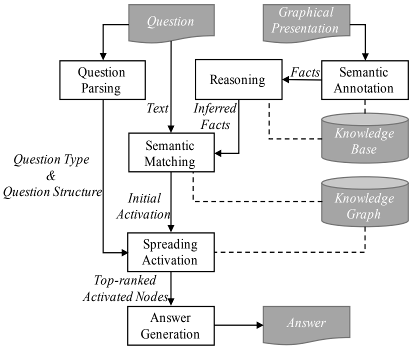

Question
Graphical Presentation
Question Parsing
Reasoning
Facts
Semantic Annotation
Text
Inferred Facts
Knowledge Base
Semantic Matching
Question Type & Question Structure
Initial Activation
Knowledge Graph
Spreading Activation
Top-ranked Activated Nodes
Answer Generation
Answer

Fig. 2: Overview of the proposed approach.

Towards meeting the identified challenges, we propose a hybrid approach which features a pipeline design and hence an interpretable problem solving process to facilitate future applications. The major design principle is to mix information retrieval techniques (e.g., ranking) with knowledge engineering, considering the sole use of information retrieval is not powerful enough and a fully formal method is unfeasible at the moment due to the complexity in knowledge representation and the difficulty in natural language understanding.

As outlined in Fig. 2, we construct a knowledge graph (KG) for the geography domain. We choose this single but flexible representation as it can help us manage the complexity of knowledge engineering. To answer a question, our approach initially activates a set of nodes in the KG that can find semantic matches in the question text, and then iteratively spreads that activation to neighboring nodes for ranking. Question parsing returns the type and structure of the question, which are used to filter out irrelevant nodes. Finally, the remaining top-ranked nodes are selected, and an answer is generated by concatenating their text.

We also plan to annotate graphical presentations and produce facts which in turn will be enriched by reasoning over a knowledge base. The resulting facts will also be considered in semantic matching. However, this branch has not been incorporated into our implementation reported in this paper.

4 Detailed Implementation

We collect a representative sample of 1,863 geography questions from Gaokao or mock exams in recent years. The current implementation of our approach is guided by an investigation of this dataset. In this section, we detail our implementation using Q1–Q3 in Fig. 1 as a running example. We also describe the gold standard we create for the output of each component of our approach on that dataset, for characterizing real questions and for evaluation.

The hybrid nature of the approach is highlighted in Section 4.4, founded on the techniques in Sections 4.1–4.3.

4.1 Question Parsing

Question Decomposition When applicable, we decompose a question into simpler atomic sub-questions to be separately solved whose answers will be concatenated in the end. An atomic question in Gaokao is normally identified by an imperative sentence starting with a particular verb. We define 44 such verbs in Chinese, e.g., *describe*, *explain*, *summarize*, and *identify*.

We also use common coordinating conjunctions (e.g., *and*) and punctuation marks representing coordination in Chinese to further decompose questions.

Gold Standard 1 *We manually analyze all the 1,863 questions, find 1,155 atomic questions, and decompose the other 708 questions (38%) into 1,556 atomic questions. There are a total of 2,711 atomic questions after decomposition.*

Table 1: Types of Questions

Causes	27.67%	Troubles	4.06%
Circumstances	9.48%	Distribution	4.02%
Effects	8.63%	Changes	3.65%
Features	7.89%	Opinions	1.77%
Comparison	7.71%	Redirection	0.48%
Measures	7.16%	Transport	0.33%
Significance	5.20%	MISC	11.95%

Table 2: Relations in the KG (Manual + Automated)

state	478+104	feature	101
effect	440	is-a	83+22
circumstance	365	factor	27
measure	257	manner	16
aspect	212+103	action	0+138
		Total	1,979+367

Question Classification and Structuring To capture the intent of a question, we detect its type and transform it into a type-specific semi-structured tuple representation of its arguments.

We identify 13 question types, plus MISC representing all the others, as shown in Table 1. We detect the type of a question using rules over 84 particular keywords in Chinese. Example rules are as follows.

- A question ending with *circumstances* is a *Circumstances* question.
- A question mentioning *hazards*, *threats*, or *risks* is a *Troubles* question.

A question will be in MISC if none of the rules is matched.

For each type of a question, we define one or several *question structures* to represent the arguments of a question. Example question structures are as follows:

$$\text{Effects}\ (\text{cause}, \text{recipient}, \text{aspect}, \text{polarity})\,, \tag{1}$$

$$\text{Circumstances}\ (\text{subject}, \text{aspect}, \text{polarity})\,. \tag{2}$$

The first structure represents a question asking the effects of a particular *polarity* (i.e., positive or negative effects, or unspecified) that the *cause* has on the *recipient*, in terms of some *aspect* which is optional. The second one asks the circumstances which the *subject* is under, of an optional *polarity*, and in terms of an optional *aspect*.

Considering that questions in Gaokao are mainly expressed in a handful of syntactic forms, we transform a question into a question structure based on template matching. An example template is as follows, written as a regular expression.

$$\begin{aligned}&((\text{In terms of})|(\text{Based on})\ \mathit{aspect},)?\\ &\text{explain}|\text{summarize the }\mathit{polarity}?\text{ effects of }\mathit{cause}\text{ on }\mathit{recipient}.\end{aligned} \tag{3}$$

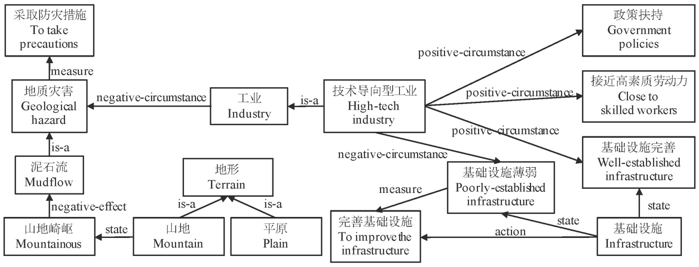

采取防灾措施
To take precautions
measure
地质灾害
Geological hazard
negative-circumstance
工业
Industry
is-a
技术导向型工业
High-tech industry
positive-circumstance
政策扶持
Government policies
positive-circumstance
接近高素质劳动力
Close to skilled workers
positive-circumstance
基础设施完善
Well-established infrastructure
is-a
泥石流
Mudflow
negative-effect
山地崎岖
Mountainous
state
地形
Terrain
is-a
is-a
山地
Mountain
平原
Plain
negative-circumstance
基础设施薄弱
Poorly-established infrastructure
measure
完善基础设施
To improve the infrastructure
action
state
state
基础设施
Infrastructure

Fig. 3: An example of knowledge graph.

Example 1. Q2a and Q3a are transformed with the question structures in Eqs. (1) and (2), respectively:

$$\begin{aligned}\text{Effects}\ (&\text{natural circumstances},\\ &\text{the economic development of Guizhou},\\ &\text{terrain}, \text{negative})\,,\end{aligned} \tag{4}$$

$$\begin{aligned}\text{Circumstances}\ (&\text{the Gui-An base for the big data industry},\\ &\text{developing}, \text{positive})\,,\end{aligned} \tag{5}$$

where the first one is obtained using the template in Eq. (3).

Gold Standard 2 *We manually classify all the 2,711 atomic questions. Their distribution is shown in Table 1. We also transform each of the 685 atomic* Circumstances*,* Effects*, and* Measures *questions into a question structure.*

4.2 Knowledge Graph Construction

Our proposed implementation focuses on answering *Circumstances*, *Effects*, and *Measures* questions as they are less dependent on graphical presentations which are not addressed in this paper. These questions constitute 25% of all.

The remaining components of our approach center around a geography knowledge graph (KG) illustrated in Fig. 3, which is semi-structured as each node is a free text that can represent a concept, a fact, or any other thing, whereas each edge represents one of a set of predefined relations. We construct the KG in a semi-automatic way. A domain expert manually constructs the first version to cover correct answers to the collected questions. The KG is then automatically completed with more edges and nodes to improve its quality.

Manual Construction The domain expert constructs the KG in four steps. First, geography concepts extracted from textbooks and the official Gaokao guide are added to the KG as nodes. Then, relations extracted from textbooks and our collected questions and answers are added as edges. These relations, namely *circumstance*, *effect*, and *measure*, provide candidate answers to the corresponding types of questions. Next, more types of relations, e.g., *is-a*, *aspect*, and *state*, are added to connect nodes. Finally, the KG is refined by merging nearly duplicate nodes and substituting some relations (e.g., *effect*) with their sub-relations having polarity information (e.g., *negative-effect*). The resulting KG consists of 1,243 nodes and 1,979 edges. Table 2 shows its distribution of relations.

Gold Standard 3 *We manually map the correct answer of each* Circumstances*,* Effects*, and* Measures *question to a set of nodes in the KG, for evaluating the output of our approach.*

Automated Completion To complement human efforts, we devise an algorithm for automatically completing the KG. Concepts mentioned in existing nodes are recognized. If a concept has no corresponding node in the KG, it will be added as a new node. The concept is connected to the nodes mentioning it via proper relations.

Specifically, concepts are recognized using predefined *POS patterns* tagged by the Stanford Parser. For example, given a node text (in Chinese) tagged with a sequence of NNs followed by VA, these NNs will identify a (new) concept node, which will be connected to the original node via *state*. In this way, 241 nodes (+19%) and 367 edges (+19%) are automatically added to the KG, as detailed in Table 2.

Example 2. From the node *Poorly-established infrastructure* which (in Chinese) is tagged with NN NN VA, *Infrastructure* (tagged with NN NN) is recognized as a (new) node, and is connected to *Poorly-established infrastructure* via *state*.

4.3 Semantic Matching

We initially activate all the nodes in the KG that are matched with the question text. A node will be activated if its content words/phrases are all matched by the question text. We realize semantic matching by expanding each content word/phrase with its synonyms and hyponyms. Synonyms are imported from an online Chinese thesaurus. Hyponyms are obtained from a hierarchy of 960 domain terms defined by a domain expert.

Example 3. The node *Terrain* in Fig. 3 is matched with the word *terrain* in Q2a. The node *High-tech industry* in Fig. 3 is matched with its hyponym *big data industry* in Q3a.

Gold Standard 4 *We manually match the text of all the* Circumstances*,* Effects*, and* Measures *questions with nodes in the KG, and label a total of 3,823 matches.*

4.4 Spreading Activation and Answer Generation

We concatenate the text of top-ranked nodes in the KG as an answer. Nodes are ranked by closeness to the initial activation found by semantic matching. For ranking, we use a spreading activation method (SA) called random walk with restart (RWR) [12]. SA iteratively spreads the amount of activation received by each node to its neighbors until convergence. During spreading, RWR allows to jump back to an initially activated node, thereby lifting the ranking of the nodes that are closer to the initial activation. This is the key step of our approach as it mixes information retrieval techniques (i.e., SA and RWR ranking) with knowledge engineering (i.e., KG) to discover implicit answers.

Example 4. As Q2a matches the node *Terrain* in Fig. 3, nodes that are closer to it will be ranked higher, e.g., *Geological hazard* compared with *Poorly-established infrastructure*.

Further, we leverage the structure of a question to filter out irrelevant nodes. For each structure we define one or several *relation path expressions* for traversing the KG, which are written as regular expressions. For example, we define the following relation path expression for the question structure in Eq. (2) when *polarity* is positive:

$$\textit{subject}\ \text{is-a}^{*}\ \text{positive-circumstance}\,, \tag{6}$$

which starts searching the KG from each node matched by the *subject* argument, passes through any number of *is-a* edges, and finally passes through a *positive-circumstance* edge and reaches a set of nodes which will be labeled as relevant; if *subject* does not match any node, search will start from every node. Irrelevant nodes will be removed from the answer. This step realizes answer type coercion with reasoning capabilities, e.g., is-a reasoning in this example.

Table 3: Percentage of Correctly Classified Questions

Causes	99.33%	Troubles	97.27%
Circumstances	98.83%	Distribution	97.25%
Effects	97.44%	Changes	100.00%
Features	93.46%	Opinions	77.08%
Comparison	93.78%	Redirection	92.31%
Measures	91.75%	Transport	77.78%
Significance	99.29%	MISC	86.42%
		Overall	95.50%

Table 4: Percentage of Correctly Structured Questions

Circumstances	88.72%
Effects	89.32%
Measures	75.26%
Overall	85.11%

Example 5. As the *subject* argument of Q3a's question structure in Eq. (5) matches the node *High-tech industry* in Fig. 3, using the relation path expression in Eq. (6) we identify three relevant nodes: *Government policies*, *Close to skilled workers*, and *Well-established infrastructure*.

The proposed problem solving process is interpretable. We can extract from the KG those paths connecting returned nodes with initially activated nodes (which are matched with the question text), and present as answer explanations.

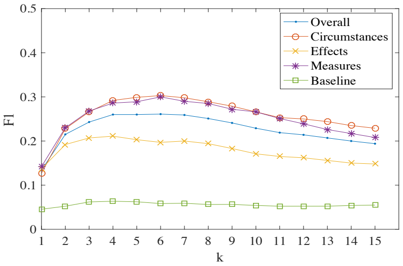


Fig. 4: F1 score of k top-ranked nodes as an answer.

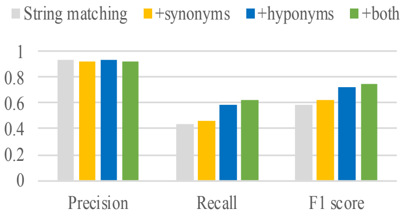


Fig. 5: Precision, recall, and F1 score of matching.

5 Evaluation

We evaluated our approach based on the collected questions and the gold standard we created. All the relevant resources have been made available at [URL to show in camera-ready].

5.1 Evaluation of the Entire Approach

Evaluation Design. In Gold Standard 3 we labeled a set of nodes in the KG as the correct answer to a question. Based on that, we configured both our approach and a baseline method to output k top-ranked nodes for each of the 685 questions, and calculated their average F1 scores on all the questions.

Baseline. We compared with BM25, a baseline widely used in information retrieval. To fit our problem, we adapted it to take nodes from *circumstance*, *effect*, and *measure* relations in the KG as candidate answers to *Circumstances*, *Effects*, and *Measures* questions, respectively, and rank nodes by the BM25 relevance between node and question text. Answers in our experiments were covered by the KG we constructed. Although semantic parsing and other knowledge-based methods for QA can work with a KG, they could not process our questions due to the unfitness discussed in Section 2.

Results. As shown in Fig. 4, the F1 score of the baseline was consistently below 0.1 under any setting of k. The overall F1 score of our approach reached 0.26 when $k = 5$, and exceeded 0.3 on *Circumstances* and *Measures* questions, significantly outperforming the baseline.

5.2 Evaluation of Components

Although largely exceeding the baseline, there was considerable room for improving our approach. We evaluated each component to identify their limitations.

Question Decomposition. Among the 1,155 indecomposable atomic questions in Gold Standard 1, our method mistakenly decomposed 40 (3%) as they contained triggering verbs or coordinating conjunctions. Among the other 708 non-atomic questions, the decomposition outputted by our method was exactly the same as the gold standard on 595 (84%); the generated sub-questions contained insignificant errors on 105 (15%), mainly due to incorrect phrase structure trees about coordinating conjunctions; the decomposition contained significant errors only on 8 (1%). To conclude, out of the 1,863 questions, the automated decomposition of 1,815 (97%) was (nearly) correct.

Question Classification and Structuring. Among the 2,711 atomic questions in Gold Standard 2, our method correctly classified 2,589 (96%). As shown in Table 3, the accuracy exceeded 90% on most question types, though relatively low on *Opinions* and *Transport* questions because they were given low priority when a question could match multiple rules leading to different types.

Among the 685 *Circumstances*, *Effects*, and *Measures* questions in Gold Standard 2, our method transformed 583 (85%) into correct question structures, as shown in Table 4. Incorrect transformation was mainly due to incorrect word segmentation, an open problem for Chinese processing.

Semantic Matching. In Gold Standard 4 we labeled 3,823 matches. Based on that, we calculated the average precision, recall, and F1 score of the matches generated by our method on all the questions. As shown in Fig. 5, the precision of string matching was high (0.94), but the recall was unsatisfactory (0.43). Expanding content words/phrases with synonyms and hyponyms, recall notably increased by 0.03 and 0.16, respectively, whereas the decreases in precision were negligible, so F1 scores rose. Using both synonyms and hyponyms, F1 score reached 0.75. Our method mainly failed to find partial matches and deep semantic matches. We did not allow partial matches because otherwise precision would drop largely. To identify deep semantic matches representing implicit relatedness, e.g., the match between the text *state-level new area* and the node *Government policies* in the KG, more sophisticated semantic matching techniques would be needed, or the KG should be substantially extended to cover more concepts and relations such that implicit matches would more likely be identified during spreading activation.

Spreading Activation and Answer Generation. In addition to the evaluation of the entire approach in Section 5.1, we also evaluated the k top-ranked nodes outputted by our spreading activation method using manually labeled matches in Gold Standard 4 as initial activation. F1 score was raised to 0.31 (from 0.26) when $k = 5$, still not very high mainly due to the incompleteness of the KG. Many concepts mentioned in question text were not covered by the KG, thereby incapable of effectively guiding the spreading activation.

Concluding Remarks. Our question parsing, though based on a few simple rules and templates, generated high-quality results. Semantic matching was satisfactory but could be improved. The incompleteness of the KG became a bottleneck of our current implementation.

6 Future Directions

We present the first research effort to answer geography questions in Gaokao. We have identified the difficulties in problem understanding and solving, and we invite the community to continue contributing solutions to this new AI challenge, as it will benefit future-generation computer-aided education and push the boundaries of AI. To provide supporting resources, we have made all our data publicly available.

Our major technical contribution is a pipeline approach which mixes information retrieval techniques with knowledge engineering and realizes interpretable problem solving. Effectively integrating question parsing, semantic matching, spreading activation and answer generation over a constructed knowledge graph, our approach exhibits a non-trivial capacity to answer complex questions and significantly outperforms a baseline in an extensive experiment. The result is promising as graphical presentations in questions and more powerful reasoning capabilities are still yet to be exploited.

Several open problems have surfaced to be addressed in future work. (i) To extend the coverage of the KG, we are seeking practically effective methods for automated construction of KG. Recent studies have shown it is possible to automatically extract specific relations of reasonably good quality from text to answer the corresponding type of questions, e.g., causality [6]. We will adapt those methods to build a large-scale KG automatically and evaluate their effectiveness. (ii) Although KG as a single flexible representation has successfully helped us manage the complexity of knowledge engineering, it is unlikely to fit all question types. We will consider and integrate other knowledge representations which are possibly associated with different types of reasoning. (iii) Despite the effectiveness of rule- and template-based question parsing in our application, more powerful and robust techniques are still needed to process user-generated questions in potential future applications which are less formal than those in Gaokao. In particular, to better understand and use more information (e.g., newly defined concepts and laws) provided in question text, we will experiment with learning [7] and semantic parsing methods [10] which have been used to solve verbally expressed number problems.

References

1. Angele, J., Moench, E., Oppermann, H., Staab, S., Wenke, D.: Ontology-based query and answering in chemistry: OntoNova @ project Halo. In: Proceedings of the Second International Semantic Web Conference. pp. 913–928. Springer-Verlag, Sanibel Island, USA (October 2003)
2. Cheng, G., Zhu, W., Wang, Z., Chen, J., Qu, Y.: Taking up the Gaokao challenge: An information retrieval approach. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. pp. 2479–2485. International Joint Conferences on Artificial Intelligence, New York, USA (July 2016)
3. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., Schlaefer, N., Welty, C.: Building Watson: An overview of the DeepQA project. AI Magazine **31**(3), 59–79 (Fall 2010)
4. 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 Ninth International Conference on Language Resources and Evaluation. pp. 2590–2597. European Language Resources Association, Reykjavik, Iceland (May 2014)
5. Gunning, D., Chaudhri, V.K., Clark, P., Barker, K., Chaw, S.Y., Greaves, M., Grosof, B., Leung, A., McDonald, D., Mishra, S., Pacheco, J., Porter, B., Spaulding, A., Tecuci, D., Tien, J.: Project Halo update - progress toward Digital Aristotle. AI Magazine **31**(3), 33–58 (Fall 2010)
6. Hidey, C., McKeown, K.: Identifying causal relations using parallel Wikipedia articles. In: Proceedings of the Fifty-Fourth Annual Meeting of the Association for Computational Linguistics. pp. 1424–1433. Association for Computational Linguistics, Berlin, Germany (August 2016)
7. Hosseini, M.J., Hajishirzi, H., Etzioni, O., Kushman, N.: Learning to solve arithmetic word problems with verb categorization. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. pp. 523–533. Association for Computational Linguistics, Doha, Qatar (October 2014)
8. Kolomiyets, O., Moens, M.F.: A survey on question answering technology from an information retrieval perspective. Information Sciences **181**(24), 5412–5434 (December 2011)
9. Liang, P.: Learning executable semantic parsers for natural language understanding. Communications of the ACM **49**(9), 68–76 (September 2016)
10. Shi, S., Wang, Y., Lin, C.Y., Liu, X., Rui, Y.: Automatically solving number word problems by semantic parsing and reasoning. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 1132–1142. Association for Computational Linguistics, Lisbon, Portugal (September 2015)
11. Tan, M., dos Santos, C., Xiang, B., Zhou, B.: Improved representation learning for question answer matching. In: Proceedings of the Fifty-Fourth Annual Meeting of the Association for Computational Linguistics. pp. 464–473. Association for Computational Linguistics, Berlin, Germany (August 2016)
12. Tong, H., Faloutsos, C.: Center-piece subgraphs: Problem definition and fast solutions. In: Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 404–413. Association for Computing Machinery, Philadelphia, USA (August 2006)
13. Yu, K., Liu, Q., Zheng, Y., Zhao, T., Zheng, D.: History question classification and representation for Chinese Gaokao. In: Proceedings of the Twentieth International Conference on Asian Language Processing. pp. 129–132. Institute of Electrical and Electronics Engineers, Tainan, Taiwan (November 2016)