GeoSQA: A Benchmark for Scenario-based Question Answering in the Geography Domain at High School Level

Zixian Huang, Yulin Shen, Xiao Li, Yuang Wei, Gong Cheng, Lin Zhou, Xinyu Dai, Yuzhong Qu
National Key Laboratory for Novel Software Technology, Nanjing University, China
{zixianhuang,ylshen,xiaoli,weiyuang}@smail.nju.edu.cn,
{gcheng,daixinyu,yzqu}@nju.edu.cn,zhoul@nlp.nju.edu.cn

Abstract

Scenario-based question answering (SQA) has attracted increasing research attention. It typically requires retrieving and integrating knowledge from multiple sources, and applying general knowledge to a specific case described by a scenario. SQA widely exists in the medical, geography, and legal domains—both in practice and in the exams. In this paper, we introduce the GeoSQA dataset. It consists of 1,981 scenarios and 4,110 multiple-choice questions in the geography domain at high school level, where diagrams (e.g., maps, charts) have been manually annotated with natural language descriptions to benefit NLP research. Benchmark results on a variety of state-of-the-art methods for question answering, textual entailment, and reading comprehension demonstrate the unique challenges presented by SQA for future research.

1 Introduction

Scenario-based question answering (SQA) is an emerging application of NLP (Lally et al., 2017). Different from traditional QA, a question in SQA is accompanied by a *scenario*, e.g., a patient summary in the medical domain asking for diagnosis or treatment. A scenario differs from a document given in the reading comprehension task where the answer can be extracted or abstracted from the document (Rajpurkar et al., 2016; Nguyen et al., 2016; Lai et al., 2017). SQA requires retrieving and integrating knowledge from multiple sources, and applying general knowledge to a specific case described by the scenario.

SQA has found application in many fields, especially in the legal domain (Ye et al., 2018; Luo et al., 2017; Zhong et al., 2018) and in high-school geography exams (Ding et al., 2018; Zhang et al., 2018). The latter is particularly challenging because a geographical scenario consists of both text and diagrams (e.g., maps, charts). Questions include city planning, climates, agriculture planning, transportation, etc. An example of a scenario and a question is presented in Figure 1.

Geographical SQA has posed great challenges to NLP and related research, ranging from scenario understanding to cross-modal knowledge integration and reasoning. However, there is a lack of large datasets and benchmarking efforts for this task. In this paper, we introduce GeoSQA—an SQA dataset in the geography domain consisting of 1,981 scenarios and 4,110 multiple-choice questions at high school level. In particular, each diagram has been manually annotated with a high-quality natural language description of its content, as illustrated in Figure 1. This labor-intensive effort significantly extends the use of GeoSQA, which can support visual SQA (using the diagrams), natural language based SQA (using the annotations of diagrams), and even the diagram-to-text research. We test the effectiveness of a variety of methods for question answering, textual entailment, and reading comprehension on GeoSQA. The results demonstrate its unique challenges, waiting for more effective solutions.

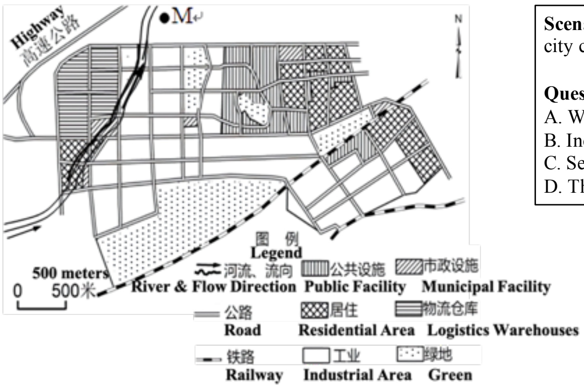
The remainder of the paper is organized as follows. Section 2 discusses related work. Section 3 describes the GeoSQA dataset. Section 4 reports benchmark results. Section 5 concludes the paper.

2 Related Work

2.1 Scenario-based Question Answering

Scenario-based question answering (SQA) is introduced by Lally et al. (2017), where the WatsonPaths system is presented to answer questions that describe a medical scenario about a patient and ask for diagnosis or treatment. SQA also finds application in the legal domain, where a legal case de-

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing
and the 9th International Joint Conference on Natural Language Processing, pages 5866–5871,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics



Scenario: The diagram shows the urban planning of a satellite city in 2010. The city commonly receives NW winds. Read the diagram and answer the question.

Question: Location M in the diagram is suitable for
A. Waterworks
B. Incineration plant
C. Sewage treatment plant **(the correct answer)**
D. Thermal power plant

Diagram annotations --- templated text:
M is to the north of the river.
M is in the northwest of the satellite city.

Diagram annotations --- free-form text:
The river flows north.

Figure 1: An example of a scenario, a question, and diagram annotations.

scribes a scenario to be decided (Ye et al., 2018; Luo et al., 2017; Zhong et al., 2018).

For some domains, reasoning with domain knowledge is essential to SQA. Therefore, such questions often appear in exams like China's version of the SAT called Gaokao. For example, for the geography domain, Ding et al. (2018) and Zhang et al. (2018) construct a knowledge graph to support answering scenario-based geography questions at high school level.

2.2 Related Datasets

There are many datasets for traditional QA, such as WebQuestions (Berant et al., 2013) and WikiQA (Yang et al., 2015). A closely related task is reading comprehension, where the answer to a question is extracted or abstracted from a given document (Rajpurkar et al., 2016; Nguyen et al., 2016; Lai et al., 2017). By comparison, SQA is arguably more difficult because a scenario is present and contextualizes a question, but no direct answer can be identified from the scenario.

The GeoSQA dataset introduced in this paper is not the first resource for geographical SQA. Ding et al. (2018) and Zhang et al. (2018) have made their datasets public. However, compared with GeoSQA, their datasets are small and, more importantly, they ignore diagrams which represent a unique challenge to geographical SQA. By contrast, diagrams are included in GeoSQA for completeness, and have been manually annotated with natural language descriptions for extended use—including but not limited to NLP research.

Existing SQA datasets for other domains include the TREC Precision Medicine track (Roberts et al., 2018) for the medical domain, and CAIL (Xiao et al., 2018) for the legal domain. However, SQA in the geography domain requires different forms of knowledge and different reasoning capabilities, and has posed different research challenges.

3 The GeoSQA Dataset

GeoSQA is an SQA dataset in the geography domain, containing 1,981 scenarios and 4,110 multiple-choice questions at high school level. A scenario consists of a piece of text and a diagram, supporting 1–5 questions. A diagram is annotated with a natural language description of its content. A question has four options that are possible answers. Exactly one of them is the correct answer. The dataset is available online[1].

3.1 Data Collection and Deduplication

We crawled over 6,000 scenarios and 13,000 questions from Gaokao and mock tests that are available on the Web. However, some scenarios are just copies or trivial variants of others. There is a need to clean and deduplicate the collected data.

Method. The problem is to decide whether a pair of scenarios are (near) duplicates or not.

We firstly establish a matching between their structures. The matching consists of 6 pairs of their text elements: 1 pair of their scenario text, 1 pair of their most similar questions, and 4 pairs of the most similar options of the above two questions. Text similarity is computed by the cosine similarity between two bags of words.

Then we extend a popular text matching method called MatchPyramid (Pang et al., 2016) to classify a pair of scenarios as duplicates or not. The original implementation of MatchPyramid can

[1] ws.nju.edu.cn/gaokao/geosqa/1.0

only process a pair of text. We extend it to process all the 6 pairs of text by concatenating their feature vectors inside MatchPyramid.

Experiments. To evaluate our method, we manually label 1,000 pairs of scenarios where positive and negative examples are balanced. The set is divided into training, validation, and test sets with a 60-20-20 split. Our method achieves an accuracy of 95.3% on the test set, showing its satisfying performance.

Then we apply our method to the entire dataset. We index all the scenarios using Apache Lucene. For each scenario, we retrieve 10 top-ranked scenarios as suspect duplicates. Each pair of scenarios is classified by our method, which is trained using all the 1,000 labeled examples.

To verify the quality of the final results, we randomly sample and manually check 100 pairs of scenarios that are predicted to be duplicates. Indeed, all of them are decided correctly. We also randomly sample 50 scenarios and, for each of them, we retrieve and manually check 10 top-ranked scenarios that are predicted to be non-duplicates. Only 6% are decided incorrectly, suggesting a low degree of redundancy of our data.

3.2 Diagram Annotation

Crawled diagrams are images. To extend the use of GeoSQA and to better support NLP research, we manually annotate each diagram with a high-quality natural language description of its content so that NLP researchers can use these text annotations instead of the original diagrams.

Annotation. We recruited 30 undergraduate students from one of the top-ranked universities in China as annotators. All of them had an excellent record in geography during high school.

Each diagram is assigned to one annotator, who also has access to the scenario text and related questions. The annotator firstly categorizes the diagram according to a hierarchy of categories. Each category is associated with a set of text templates that are recommended to be used in annotations as far as possible. However, the annotator is free to use any form of text to annotate information that is not covered by the provided templates.

Annotations are required to precisely reflect the content of the diagram. All the information related to every supported question and every option should be annotated. On the other hand, inferring new knowledge via human reasoning is prohibited.

Note that the entire annotation process is designed to be iterative. The 22 diagram categories and 81 text templates are not predefined but incrementally induced during the experiment. However, there are still 11% of the diagrams that are believed to not belong to any category. No templates are provided for their annotations.

An example of annotations is shown in Figure 1.

Audit. To ensure the quality of the annotations, we recruited 3 senior annotators to audit the results. Each diagram is audited by one senior annotator, who rates the annotations from three dimensions in the range of 1–5.

- Sufficiency: The annotations cover all the necessary information in the diagram that is useful for answering related questions.
- Fairness: The annotations are not biased towards any particular option of a question.
- Objectiveness: The annotations are plain descriptions of the diagram—not influenced by human reasoning.

The scenarios where the annotations of the diagram are rated below 3 in any dimension are excluded from the dataset.

4 Benchmark Results

We tested several state-of-the-art methods for question answering, textual entailment, and reading comprehension on our GeoSQA dataset.

4.1 Corpora

We use two corpora as background knowledge. **Textbooks** contains 15K sentences extracted from two high-school geography textbooks. **Wikipedia** contains 1M articles in the latest Chinese edition of Wikipedia. We index their sentences using Apache Lucene.

4.2 Methods

We tested two text matching methods. In **IR** (Clark et al., 2016), for each option, we use a combination of the scenario text, the question, and the option as a query, to retrieve the top-ranked sentence from a corpus. We use the ranking score of this sentence as the score of the option. Finally, we choose the option with the highest score as the answer. In **PMI** (Clark et al., 2016), for each option, we calculate the Pointwise Mutual Information (PMI) between the question and the option as the score of the option. Finally, we choose the option with the highest score as the answer. Probabilities in PMI are estimated based on a corpus.

We tested four textual entailment methods: **ESIM** (Chen et al., 2017), **DIIN** (Gong et al., 2018), **BERT**$_{NLI}$ (Devlin et al., 2018), and **BiMPM** (Wang et al., 2017). The first three methods were trained on the XNLI dataset (Conneau et al., 2018). The last method was trained on the LCQMC dataset (Liu et al., 2018). For each option, a textual entailment method retrieves six top-ranked sentences from a corpus to form the entailing text. Retrieval follows the procedure described in the above-mentioned **IR** method. The scenario text and diagram annotations may or may not be included in the entailing text, depending on the configuration. A combination of the question and the option form the entailed text. Finally, we choose the option with the highest entailment score as the answer.

We tested one reading comprehension method: **BERT**$_{RC}$ (Devlin et al., 2018). It was trained on the DuReader dataset (He et al., 2018). For each option, a reading comprehension method retrieves six top-ranked sentences from a corpus as part of the passage for reading comprehension. Retrieval follows the procedure described in the above-mentioned **IR** method. The scenario text and diagram annotations may or may not be included in the passage, depending on the configuration. Finally, the reading comprehension method extracts a text span from the passage. We choose the option that is the most similar to the extracted text span as the answer. Similarity is computed by the cosine similarity between two bags of words.

4.3 Results

The results are summarized in Table 1. Note that a question has four options. Even guessing randomly, the expected proportion of correctly answered questions would be 25%.

Almost all the methods performed similar to random guess, showing that SQA on our dataset has its unique challenges.

	Textbooks	Wikipedia
IR	25.24	25.14
PMI	26.22	25.19
ESIM w/o scenario	25.85	25.41
ESIM w/ scenario	24.34	24.41
DIIN w/o scenario	24.15	25.20
DIIN w/ scenario	25.11	24.89
BERT$_{NLI}$ w/o scenario	24.29	24.17
BERT$_{NLI}$ w/ scenario	24.97	24.68
BiMPM w/o scenario	24.13	24.51
BiMPM w/ scenario	24.76	23.81
BERT$_{RC}$ w/o scenario	24.81	24.78
BERT$_{RC}$ w/ scenario	23.66	23.01

Table 1: Proportions of correctly answered questions.

4.4 Discussion

To explain the poor performance of existing methods, we have identified the following challenges. First, SQA relies on domain knowledge that is not provided in the scenario. However, relevant knowledge may fail to be retrieved from the corpus. Second, for some questions, commonsense knowledge is needed but is not included in textbooks and may fail to be retrieved from Wikipedia. Third, the retrieved general knowledge needs to be applied to the specific case described by a scenario. Existing QA and reading comprehension methods hardly have this capability.

5 Conclusion

We have contributed GeoSQA—a large SQA dataset where diagrams are present and have been manually annotated with natural language descriptions. We have tested a variety of existing methods on our dataset. The results are not satisfactory, thus demonstrating the unique challenges presented by the SQA task on our dataset. In future work, we will work towards more effective solutions to meet the challenges.

Researchers are invited to use GeoSQA to support their own tasks, including but not limited to natural language based SQA, visual SQA, and the diagram-to-text task.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1005100, and in part by the NSFC under Grant 61572247. Gong Cheng was supported by the Qing Lan Program of Jiangsu Province. We would like to thank all the annotators.

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013*

Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL, pages 1533–1544.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1657–1668.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter D. Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 2580–2586.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2475–2485.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jiwei Ding, Yuan Wang, Wei Hu, Linfeng Shi, and Yuzhong Qu. 2018. Answering multiple-choice questions in geographical gaokao with a concept graph. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, pages 161–176.

Yichen Gong, Heng Luo, and Jian Zhang. 2018. Natural language inference over interaction space. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. Dureader: a chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering@ACL 2018, Melbourne, Australia, July 19, 2018*, pages 37–46.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 785–794.

Adam Lally, Sugato Bagchi, Michael Barborak, David W. Buchanan, Jennifer Chu-Carroll, David A. Ferrucci, Michael R. Glass, Aditya Kalyanpur, Erik T. Mueller, J. William Murdock, Siddharth Patwardhan, and John M. Prager. 2017. Watson-Paths: Scenario-based question answering and inference over unstructured information. *AI Magazine*, 38(2):59–76.

Xin Liu, Qingcai Chen, Chong Deng, Huajun Zeng, Jing Chen, Dongfang Li, and Buzhou Tang. 2018. LCQMC: A large-scale chinese question matching corpus. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 1952–1962.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2727–2736.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016.*

Liang Pang, Yanyan Lan, Jiafeng Guo, Jun Xu, Shengxian Wan, and Xueqi Cheng. 2016. Text matching as image recognition. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 2793–2799.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392.

Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, William R. Hersh, Steven Bedrick, and Alexander J. Lazar. 2018. Overview of the trec 2018 precision medicine track. In *Proceedings of The Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018.*

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4144–4150.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A large-scale legal dataset for judgment prediction. *CoRR*, abs/1807.02478.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 2013–2018.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1854–1864.

Zhiwei Zhang, Lingling Zhang, Hao Zhang, Weizhuo He, Zequn Sun, Gong Cheng, Qizhi Liu, Xinyu Dai, and Yuzhong Qu. 2018. Towards answering geography questions in gaokao: A hybrid approach. In *Proceedings of the Third China Conference on Knowledge Graph and Semantic Computing, CCKS 2018, Tianjin, China, 14-17 August 2018*, pages 1–13.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3540–3549.