GCRC: A New MRC Dataset from Gaokao Chinese for Explainable Evaluation

Hongye Tan[1*], **Xiaoyue Wang**[1*], **Yu Ji**[1], **Ru Li**[†], **Xiaoli Li**[2], **Zhiwei Hu**[1], **Yunxiao Zhao**[1] **and Xiaoqi Han**[1]
1.School of Computer and Information Technology, Shanxi University, Taiyuan, China
2.Institute for Infocomm Research, A*Star, Singapore
{wangxy0808, jiyu0515, yunxiaomr, xiaoqisev}@163.com
{liru,tanhongye}@sxu.edu.cn
xlli@ntu.edu.sg, zhiweihu@whu.edu.cn

Abstract

Recently, driven by numerous publicly available machine reading comprehension (MRC) datasets, MRC systems have made some progress. These datasets, however, have two major limitations: 1) the defined tasks are relatively simple, and 2) they do not provide explainable evaluation which is critical to objectively and comprehensively review the reasoning capabilities of current MRC systems. In this paper, we propose GCRC, a new dataset with challenging and high-quality multi-choice questions, collected from Gaokao Chinese (Chinese subject from the National College Entrance Examination of China). We have manually labelled three types of evidence to evaluate MRC systems' reasoning process: 1) sentence-level relevant supporting facts in an article required for answering a given question, 2) error reason of a distractor (i.e., an incorrect option) for explaining why a distractor should be eliminated, which is an important reasoning step for multi-choice questions, and 3) types of reasoning skills required for answering questions. Extensive experiments show that our proposed dataset is more challenging and very useful for identifying the limitations of existing MRC systems in an explainable way, facilitating researchers to develop novel machine learning and reasoning approaches to tackle this challenging research problem.[1]

1 Introduction

Machine Reading Comprehension (MRC) is a critical task in many real-world applications, which requires machines to understand a text passage, and answer relevant questions. It evaluates machines' understanding and reasoning capabilities on the underlying natural language text. Numerous MRC datasets have been proposed and facilitate the progress of MRC systems, which have achieved near-human performance on some datasets. However, this does not indicate the MRC systems have owned human-like language understanding and reasoning capabilities.

One reason is that questions in current data are not challenging enough, leading to most of them get "solved" very soon. (Sugawara et al., 2018) demonstrate that in many MRC datasets, a considerable number of easy questions can be answered based on the first few tokens of the questions or word matching, without complex reasoning capabilities. The other reason is that most datasets only provide a black-box and overall evaluation on the accuracy of predicted answers, which does not provide **explainable evaluation** on a system's internal reasoning capabilities. In other words, it is unable to explain whether a system gets a correct answer via the right reasoning process, and not enough to identify the specific limitations of a system.

Recently, in order to address the problems mentioned, some datasets, focusing on reasonings and providing additional information to evaluate the internal reasoning steps of a system, have been proposed. For example, MultiRC (Khashabi et al., 2018) and HotpotQA (Yang et al., 2018) include the questions requiring multi-sentence reasoning

[*]These authors contributed equally.
[†]Corresponding author
[1]Resources will be available through https://github.com/SXUNLP/GCRC

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1319–1330

	In Chinese	In English
Article	随着全球人口的不断增长和科学技术的飞速发展，人类在创造文明的同时也缔造了一个深受人类影响的全球生态系统。长期以来对生物资源及土地的过度利用，导致了动植物栖息地丧失、环境污染等一系列问题的出现，生态环境及生物系统遭受了严重破坏。据专家估计，由于人类活动和气候变化，地球上的生物种类目前正在以相当于正常水平1000倍的速度消失，全球已有约3.4万种植物和5200多种动物濒临灭绝，物种分布发生了大范围的变化，这些形成了全球性的生物多样性危机。... 生物多样性为人类发展带来了巨大财富，目前它却面临着来自城市化等方面的威胁。城市化对生物多样性的影响成为生态学研究者关注的焦点问题。	With the continuous growth of global population and rapid development of science and technology, human beings has created a global ecosystem which is deeply influenced by mankind while creating civilization. The long-term over-utilization of biological resources and land has led to the emergence of problems such as the loss of wildlife habitats and environmental pollution, and meanwhile, the ecological environment and biological systems have been severely damaged. According to experts, the species on the earth are disappearing at a rate of 1,000 times the normal level due to human activities and climate change. Currently, about 34,000 species of plants and more than 5,200 species of animals are on the verge of extinction, and the species distribution has been changed significantly, which has resulted in a global biodiversity crisis. ... Biodiversity has brought huge wealth to human development, but it is facing the threat from urbanization and other aspects. The impact of urbanization on biodiversity has become the focus of ecological researchers.
Question stem	根据材料一，下列理解和分析，符合文意的一项是?	Which of the following statements agrees with the information given in the article?
Options	A：深受人类影响的全球生态系统利于缓解生物多样性危机。 （所需推理能力：归纳推理） （错误原因：主谓宾关系错误） *B：第一段通过列举数据来凸显生物多样性危机的严重程度。 （所需推理能力：鉴赏分析） C：生态学者关注的焦点是生物多样性危机给人类带来哪些损失。 （所需推理能力：细节推理） （错误原因：细节错误） D：这则材料反映了对生物多样性危机的担忧并提出了应对策略。 （所需推理能力：归纳推理） （错误原因：无中生有）	A: The global ecosystem, which is deeply influenced by human beings, helps alleviate the biodiversity crisis. (Reasoning skill: Inductive reasoning) (Error reason: Wrong subject-predicate-object relationship) *B: *The first paragraph* highlights the severity of the biodiversity crisis by giving some statistics. (Reasoning skill: Appreciative analysis) C: The focus of ecological researchers is what loss the biodiversity crisis has brought to human beings. (Reasoning skill: Detail understanding) (Error reason: Wrong details) D: *This article* reflects the concerns about biodiversity crisis and proposes countermeasures. (Reasoning skill: Inductive reasoning) (Error reason: Irrelevant to the article.)

Figure 1: An annotated instance in GCRC, which is from the real Gaokao Chinese in 2019 (* indicates the correct option). The sentences marked in yellow/green are respectively the SFs for Option A and C. As mentioned in Option B, the SFs of Option B are the first paragraph. Similarly, the SFs of Option D are the whole article. The contents marked in blue are the required reasoning skills for the options and the ERs of distractors.

or multi-hop reasoning. Moreover, MultiRC and HotpotQA both provide *sentence-level supporting facts (SFs)*, which can be used as a kind of internal explanation for the answers, although locating SFs is just the first step for question answering (QA), as many questions need to integrate the SF information with reasoning. R4C (Inoue et al., 2020) and 2WikiMultiHopQA (Ho et al., 2020) introduce *a chain of facts* (reflecting entity relationships) as the derivation steps to evaluate the internal reasoning step of systems.

However, these kinds of reasonings are quite limited and not enough to answer those complex and comprehensive questions. For example, Figure1 shows a multi-choice question, where the option D "***This article reflects*** *the concerns about biodiversity crisis and proposes countermeasures*" is a summarization of the whole given article, and its judgement cannot be made by simply reasoning over some entity relationships. Instead, the reasoning over the full information across the article will be needed.

To address the above real challenges, we propose GCRC, a new challenging dataset with 8,719 multi-choice questions, collected from reading comprehension (RC) tasks of Gaokao Chinese (short for Chinese subject from the National College Entrance Examination of China). GCRC is of high quality and high difficulty level, because Gaokao Chinese examinations are designed by educational experts, with high-level comprehensive questions and complicated articles, aiming to test the language comprehension of adult examinees.

In order to provide **explainable evaluation**, instances (e.g. Figure 1) in the dev. and test sets (including 1725 questions and 6900 options, as presented in Table 2 and Section 4) of GCRC are annotated with three kinds of information: (1) ***sentence-***

level supporting facts **(SFs)**, serves as the basic evaluation for a system's internal reasoning; and (2) ***error reasons*** **(ERs)** ***of a distractor*** (i.e. an incorrect option) for explaining why a distractor should be eliminated. In Gaokao Chinese, distractors are often designed in a very confusing way and look like a correct option, although they are incorrect. Identifying the semantic difference between a distractor and the given article, and knowing exactly why a distractor is wrong could help systems or examinees to choose the correct answers. Therefore, we spend considerable effort to thoroughly understand articles and manually label seven types of ERs of a distractor as a form of internal reasoning step for multi-choice questions; and (3) ***types of reasoning skills*** required for answering questions in Gaokao Chinese. We introduce eight typical reasoning skills in GCRC, which enable us to evaluate whether a system has corresponding reasoning capability at individual question level.

Our main contributions can be summarized as:

- We construct a new challenging dataset GCRC that is collected from Gaokao Chinese, consisting of 8,719 multiple-choice questions, which will be released on Github in future.
- Three kinds of critical information are manually annotated for explainable evaluation. To the best of our knowledge, GCRC is the first Chinese MRC dataset to provide most comprehensive, challenging, high-quality articles and questions for more deep explainable evaluation on different MRC system performance. In particular, error reasons of a distractor are the first introduced in this area as an important reasoning step for complex multi-choice questions.
- Extensive experiments show that GCRC is not only a more challenging benchmark to facilitate researchers to develop novel models, but also help us identify the limitations of existing MRC systems in an explainable way.

2 Related Work

2.1 Datasets from standard tests

There exist some datasets, collected from standard exams/tests, including English datasets RACE (Lai et al., 2017), DREAM (Sun et al., 2019), ARC (Clark et al., 2018), ReClor (Yu et al., 2020), etc., and Chinese datasets C3 (Sun et al., 2020), MCQA (Guo et al., 2017), GeoSQA (Huang et al., 2019) and MedQA (Zhang et al., 2018) etc.

Some of these datasets are of specific domains. For example, ARC is of the science domain and the questions are from the American science exams from 3rd to 9th grade. ReClor targets logical reasoning and the questions are collected from the Law School Admission Council. MCQA and GeoSQA are extracted from Gaokao History and Gaokao Geography respectively. Other datasets cover generic topics. For example, RACE and DREAM are both collected from the English exams for Chinese students, while C3 is collected from the Chinese exams for Chinese-as-a-second-language learners.

GCRC is more similar to C3, RACE and DREAM, as their questions are all generic ones. However, their question difficulty level is different, because questions in GCRC target for adult native speakers, while questions in other datasets target for second-language learners. As such, the GCRC is much more challenging than other datasets. Additionally, GCRC provides three kinds of rich information for more deep explainable evaluation.

2.2 Datasets with explanations

Some datasets provide explanation information. (Wiegreffe and Marasovic, 2021) identified three types of explanations: highlights, free-text and structured explanations. Highlights are subsets of the input elements (words, phrases, snippets or full sentences) to explain the prediction. Free-text explanations are textual or natural language explanations containing the information beyond the given input. Structured explanations have various forms for specific datasets. One of the most common form is a chain of facts as the derivation of multi-hop reasoning for an answer.

(Inoue et al., 2020) classified explanations into two types: justification explanation (collections of SFs for a decision) and introspective explanation (a derivation for a decision), which are respectively corresponding to highlights and structured explanations.

For the MRC task, a few datasets are with explanation information. For example, MultiRC (Khashabi et al., 2018) and HotpotQA (Yang et al., 2018) provides sentence-level SFs, belonging to justification explanations, to evaluate a system's

ability to identify SFs from articles that related to a question/option. R4C (Inoue et al., 2020) and 2WikiMultiHopQA (Ho et al., 2020) provide both justification and introspective explanations. In both datasets, introspective explanations are a set of factual entity relationships. Their difference is that the explanation of R4C is of semi-structured form, while the explanation of 2WikiMultiHopQA is of structured data form.

The dataset C3 provides the types of required prior knowledge for 600 questions, and its goal is to study how to leverage various knowledge to improve the system's ability for text comprehension.

Inspired by these datasets, we provide more rich explanations in GCRC. Different from the existing work, besides SFs, we provide two additional information. Specifically, we propose innovative ERs of a distractor as a reasoning step for multi-choice questions. In addition, we annotate the required reasoning skills for each instance, which enable us to clearly identify the limitations of existing MRC systems. Note that prior knowledge annotated in C3 is different from reasoning skills in GCRC, as prior knowledge in C3 mainly include linguistic, domain specific, and general world knowledge, while reasoning skills in GCRC focus on the abilities of making an inference. Generally, reasoning needs to integrate prior knowledge and the information in the given text. As far as we know, *GCRC is the first Chinese MRC dataset with rich explainable information* for evaluation purpose.

3 Dataset Overview

Questions in GCRC are of multi-choice style. Specifically, given an article D, a question Q and its options O, the evaluation based on GCRC will measure a system from the following aspects:

- *QA accuracy*. It is a common evaluation metric, also provided by the other QA datasets.
- *The performance of locating the SFs in D*, which is a basic evaluation metric to assess whether MRC system can collect all necessary sentences from articles before reasoning.
- *The performance of identifying ERs of a distractor*, which evaluates whether a system is able to exclude incorrect options by deep reasoning for multi-choice questions.
- *The performance of different reasoning skills required by questions*. This evaluation shows the limitations of reasoning skills for a MRC system.

Next, we clearly define the ERs of a distractor and the reasoning skills required for choosing a correct option in Section 3.1 and Section 3.2 respectively. The ERs of a distractor are concrete and focus on the forms of the errors, while reasoning skills are more abstract, referring to the abilities of making an inference.

3.1 Error reasons of a distractor

As mentioned in Section 1, knowing exactly why a distractor is wrong will help MRC systems or an examinee select correct answers. Therefore, we introduce error reasons of a distractor as an important internal reasoning step, and present seven typical ERs of a distractor after investigating the instances in GCRC.

Wrong details. The distractor is *lexically similar to the original text, but has a different meaning caused by alterations of some words*, such as modifiers or qualifier Words.

Wrong temporal properties. The distractor describes an events with *wrong temporal properties*. Generally, we consider five temporal properties defined in (Zhou et al., 2019): duration, temporal ordering, typical time, frequency, and stationary.

Wrong subject-predicate-object triple relationship. The distractor has a triple relationship of subject-predicate-object (sub-pred-obj) , but *the relationship conflicts with the ground truth triple in the given article*, caused by substituting one of the components in the triple. For example, in Figure 1, Option A "*The global ecosystem, which is deeply influenced by human beings, helps alleviate the biodiversity crisis.*" has a sub-pred-obj triple *"the global ecosystem—**alleviate**—the biodiversity crisis*. However, from the given article, the correct triple is *"the global ecosystem—**result in** —the biodiversity crisis"*, leading to the wrong distractor.

Wrong necessary and sufficient conditions. The distractor *intentionally misinterprets the necessary and sufficient conditions* expressed in the original article. A necessary condition is a condition that must be present for an event to occur, while a sufficient condition is a condition or set of conditions that will produce the event. For example, the statement "*Plants will grow as long as*

with air" is wrong, as air is the necessary condition for plant growing, but besides air, plants also need water and sunshine to grow.

Wrong causality. The distractor reverses/confuses the causes and effects mentioned in an original article or add non-existent causality. Generally, the cause-effect relationship is explicitly expressed with causal connectives in distractors.

Irrelevant to *the question*. The distractor's contents are indeed mentioned in the given article, but are irrelevant to the current question.

Irrelevant to *the article*. The distractor's contents are not mentioned in the original article. For example, in Figure 1, "*proposes countermeasures*" expressed in Option D "*This article reflects the concerns about biodiversity crisis and proposes countermeasures*." is not mentioned in the given article. Thus, Option D should be excluded.

3.2 Required reasoning skills

In Gaokao Chinese, RC tasks measure an examinee's text understanding and logical reasoning abilities from different perspectives. We investigate the skills required for answering questions in GCRC, and introduce eight typical skills, which are organized into the following three levels according to the amount of information needed and the complexity of reasoning for QA. Some of the reasoning skills (marked with *) are similar to those proposed by Sugawara et al. (2017).

3.2.1 Level 1: Basic information capturing

This level covers the ability to capture relevant detailed information distributed across the article, and combine them to match with options.

Detail understanding. It focuses on distinguishing the semantic differences between the given article and an option. The option, most of the time, preserves the most lexical surface form of the original article, but has some minor differences in details by using different modifiers or qualifying words.

3.2.2 Level 2: Local information integration

This level covers how to identify different types of relationships linked between sentences in an article. In Gaokao Chinese, the relationship's expressions are often implicit in a given article.

Temporal/spatial reasoning*. It aims to understand various temporal or spatial properties of events, entities and states.

Coreference resolution*. It aims to understand the coreference and anaphoric chains by recognizing the expressions referring to the same entity in the given article.

Deductive reasoning. It focus on taking a general rule or key idea described in an article, and applying it to make inferences about a specific example or phenomenon expressed in the option. For example[2], the statement of "*we rely on mobile navigation to travel and lose the ability to identify routes.*" is a specific phenomenon of the statement of "*while we are training artificial intelligence systems, we may also be trained by artificial intelligence systems*".

Mathematical reasoning*. It performs some mathematical operations, such as numerical sorting and comparison, to obtain a correct option.

Cause-effect comprehension*. It aims to understand the causal relationships explicitly or implicitly expressed in a given article.

3.2.3 Level 3: Global information integration

This level involves information integration of multiple sentences or the whole articles to comprehend the main ideas, article organization structures and the authors' emotion and sentiment.

Inductive reasoning. It integrates information from separate words and sentences, and makes inferences about an option, which is often a summarization of several sentences, a paragraph or the whole article.

Appreciative analysis. It aims to understand the article organization method, the authors' writing style and method, attitude, opinion and emotional state. For example, in Figure 1, Option B "*The first paragraph highlights the severity of the biodiversity crisis by giving some statistics*" is the analysis result of authors' writing style and method.

4 Construction and Annotation of GCRC

We have spent tremendous effort to construct the important GCRC dataset. Firstly, we search and download about 10,000 latest (year 2015 to 2020)

[2]This example is translated from real Gaokao Chinese in 2018.

multiple-choice questions of real and mock Gaokao Chinese from five websites. Then, for preprocessing, we remove those duplicated questions and questions with figures and tables, and keep those questions with four options (only one of them is correct). Next, we identify and rectify some mistakes, such as typos, in articles or corresponding questions/options. Finally, the total number of questions in GCRC is 8,719. Moreover, we adjust the labels of the correct answers, making them evenly distributed over the four different options, i.e., A, B, C, D.

Next, we will annotate the questions according to the following three steps. We would emphasize that the annotation process is extremely challenging, as it requires human annotators to fully understand both syntactic structure and semantic information of the articles and corresponding questions and options, as well as identify the error reasons of distractors, and reasoning skills required.

- **Step 1: Annotation preparation.** We first prepare an annotation guideline, including task definition and annotated examples. Then, we invite 12 graduate students of our team to participate in the annotation work. To maintain high quality and consistent annotations, our annotators first annotate these questions individually, and subsequently discuss and reach agreements if there is any discrepancy between two annotators. The process further improves the annotation guideline and better trains our annotators.
- **Step 2: Initial annotations.** Firstly, each question is annotated by an annotator independently, where we show an article, question, all candidate options and the label of the answer. Annotators have completed the following three tasks: (1) select the sentences in the article, which are needed for reasoning, as the sentence-level SFs; (2) provide ERs of a distractor from the types discussed in Section 3.1; (3) provide the reasoning skills required for each option based on the skills described in Section 3.2.
- **Step 3: Annotation consistency.** When two annotators disagree with their own annotations, we invite the third annotator to discuss with them, and reach the final annotations. In the rare cases where they cannot agree with each other, we will keep the annotations with at least two supports.

The annotators' consistency is evaluated by the Inter Annotator Agreement IAA) value and the IAA value is 83.8%.

As mentioned before, manual annotations of GCRC is expensive and complicated, because questions in Gaokao Chinese are designed for adult native speakers and thus are challenging. It requires deep language comprehension to solve these questions. As a result, making explainable annotations in GCRC across multiple levels implies *GCRC is a very precious dataset with valuable annotations*. Although the size of the dataset is not too big, it is big enough to be used for providing diagnosis of existing MRC systems. It also provides an ideal testbed for researchers to propose novel transfer learning or few-shot learning methods to solve the tasks.

Statistics. We partition our data into training (80%), development (dev, 10%) and test set (10%), mainly according to the number of questions. We have annotated three types of information for a subset of GCRC. Specifically, we annotate the sentence-level SFs for 8,084 options (of 2,021 questions) sampled from the training set, and 6,900 options (of 1725 questions) in the dev and test sets.

In addition, we annotate ERs of 6,159 distractors in the training, dev and test sets, and reasoning skills for 6,900 options in the dev and test sets. We believe our dataset with relatively big annotation sizes can ensure us to identify the limitations of existing RC systems in an explainable way. Table 1 shows the the details of GCRC data splitting and corresponding annotation size.

Table 2 shows the detailed comparisons for GCRC and other three RC datasets collected from standard exams, including C3 (Sun et al., 2020), RACE (Lai et al., 2017), DREAM (Sun et al., 2019). As shown in Table 2, we observe that GCRC is the longest in terms of the average length of articles, questions and options.

Table 3 shows the distribution of types of reasoning skills based on the dev and test sets. We observe that 48.77% questions need detail understanding, and 33.10% questions require inductive reasoning. In addition, Figure 2 presents the ERs types distribution of distrators based on the dev and test sets, in which 33.2% of distractors are with wrong details and 26.4% of distractors include information irrelevant to the corresponding articles.

Splitting	Train	Dev	Test	Total
# of articles	3790	683	620	5093
# of questions	6994	863	862	8719
# of questions/options with SFs	2021/8084	863/3452	862/3448	3746/14984
# of questions/distractors with ERs	2000/3261	863/1428	862/1470	3725/6159
# of questions/options with rea. ski.	-	863/3452	862/3448	1725/6900

Table 1: Statistics of data splitting and annotation size.

	GCRC	C3	GCRC	C3	RACE	DREAM
	(in Chinese Characters)		(in tokens)			
Article len. (avg)	1119.2	116.9	329.6	53.8	352.4	66.1
Question len. (avg)	20.9	12.2	14.2	7.8	11.3	7.4
Option len. (avg)	43.8	5.5	27.1	3.2	6.7	4.2

Table 2: Statistics and comparison among four MRC datasets collected from standard exams.

5 Experiments

With our newly constructed GCRC dataset, it is interesting to evaluate the performance of existing models, and better understand GCRC's characteristics comparing with other existing data sets.

5.1 QA accuracy and GCRC difficulty level

We evaluate the QA performance of several popular MRC systems on GCRC, which will reflect the difficulty level of GCRC. Specifically, for comprehensive evaluation, we employ four models, including one *rule-based model* and three recent *state-of-the-art models based on neural networks*.

- **Sliding window** (Richardson et al., 2013). It is a rule-based baseline and chooses the answer with the highest matching score. In particular, it has TFIDF style representation and calculates the lexical similarity between a sentence (via concatenating a question and one of its candidate options) and each span in the given article with a fixed window size.

- **Co-Matching** (Wang et al., 2018). It is a Bi-LSTM-based model and consists of a co-matching component and a hierarchical aggregation component. The model not only matches the article with a question and each candidate option at the word-level, but also captures sentence structures of the article. It has achieved promising results on RACE.

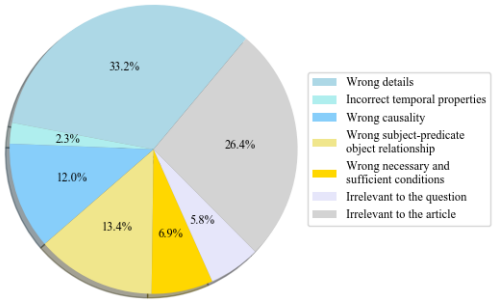

Figure 2: Distirbution(%) of types of ERs of distractors based on the dev and test sets, including 3,725 questions and 6,159 distractors.

	Reason skills	Dev	Test	all
Level 1: Basic information capturing	Detail	47.11	50.44	48.77
Level 2: Local information integration	Temporal/Spatial	0.84	0.73	0.79
	Co-reference	5.42	4.88	5.15
	Deductive	3.53	3.79	3.66
	Mathematical	0.45	0.67	0.56
	Cause-effect	5.67	4.76	5.21
Level 3: Global information integration	Inductive	34.75	31.45	33.10
	Appreciative	2.23	3.30	2.76
Leve 1: Basic information capturing		47.11	50.44	48.77
Level 2: Local information integration		15.91	14.83	15.36
Level 3: Global information integration		36.98	34.75	35.87

Table 3: Distribution (%) of types of required skills based on the annotated examples, totally including 1,725 questions and 6,900 options.

Note that we have used 300-dimensional word embeddings based on GloVe (Global Vectors for Word Representation).[3]

- **BERT** (Devlin et al., 2019). It is a pre-trained language model, which adopts multiple bidirectional transformer layers and self-attention mechanisms to learn contextual relations between words in a text. BERT has achieved very good performance on numerous NLP tasks, including the RC task defined by SQuAD. We employ Chinese BERT-base and English BERT-base released on the website.[4]

- **XLNet** (Yang et al., 2019). It is a generalized autoregressive pretraining model, which uses a permutation language modeling objective to combine the advantages of autoregressive and autoencoding methods. XLNet outperforms BERT on 20 tasks. For experiments on Chinese datasets, we use XLNet-large and its Chinese version released on the website.[5]

[3]The English word embedding:https://nlp.stanford.edu/projects/glove/; and the Chinese word embedding: https://nlp.stanford.edu/projects/glove/

[4]https://github.com/google-research/bert

[5]https://github.com/brightmart/xlnet_zh

In addition to evaluate the performance of different models on GCRC, we also want to see how these models perform on other datasets, namely C3, RACE, and DREAM. In order to fairly and objectively compare with them, we randomly sample from these datasets, and create three new datasets with the same sizes of the splits (train, dev, test) as GCRC (shown in Table 1). The hyper-parameters of these neural baselines can be seen in Appendix.

Human Performance. We obtain the human performance on 200 questions, which are randomly sampled from the GCRC test set. We invite 20 high school students to answer these questions, where they are provided with the questions, options and articles. The average accuracy of human is 83.18%.

We first use the training sets from four datasets to train four models, which are subsequently used to evaluate their accuracies on respective test data. Dev sets are used to tune the values of the parameters of different models. Table 4 shows the comparison results. We observe that the neural network based models outperform the rule based sliding window model. In addition, pretrained models perform better than the non-pretrained models in most cases. Moreover, it is interesting to observe that all four existing models generally perform worse on GCRC than other three datasets, and the human performance on GCRC is also the lowest among the four datasets. This clearly indicates that GCRC is more challenging. Meanwhile, it can be seen that the performance gap between human and the best system on GCRC is 46.45%, suggesting that there is sufficient room for improvement, which contradicts with some overly optimistic claims that machines have exceeded humans on QA. As there is a significant performance gap between machines and humans, GCRC dataset facilitates researchers to develop novel learning approaches to better understand its questions and bridge the huge gap.

	GCRC	C3	RACE	DREAM
Sliding window	27.70%	36.70%	30.85%	40.08%
Co-Matching	35.73%	45.01%	35.38%	48.91%
BERT	30.74%	64.96%	46.13%	53.36%
XLNet	35.15%	60.90%	50.00%	59.63%
Human	83.18%	96%*	95.5%*	94.5%*

Table 4: Accuracy comparison between four computational models and human on four benchmark datasets (* indicates the performance is based on the annotated test set and copied from the corresponding paper).

In the next few subsections, we will study whether a representative model, i.e. BERT, can perform well in the explainable evaluation related subtasks, namely locating sentence-level SFs, identifying ERs, and the performance of different reasoning skills required by questions.

5.2 Performance of locating sentence-level SFs using BERT

We investigate whether BERT benefits from sentence-level SFs. We conduct an experiment, in which we input the ground-truth SFs, instead of the given article, to BERT for question answering. We observe that the accuracy of QA on the test set of GCRC is 37.47%. Comparing with the result shown in Table 4, the accuracy is increased by 6.73%, indicating that locating SFs is helpful for QA. On the other hand, we can see that the improvement is still far from closing the gap to human performance because the questions are difficult and need further reasoning to solve.

Due to the usefulness of SFs, we train a BERT model to identify whether a sentence belongs to SFs. We regard this task as a sentence pair (i.e., an original sentence in the given article and an option) classification task. We use the GCRC subset (2,021/863/862 questions as shown in Table 1) annotated with SFs for training and testing, where the training, dev and test sets include 200,798/85,024/83,656 sentence-option pairs respectively. The experimental results are shown in Table 5. We can see that performance of locating the SFs is still low (in terms of precision P, recall R and F1 measure), indicating that accurately obtaining the SFs in GCRC is challenging and directly applying BERT will not work well.

	P	R	F1
Dev	78.07%	50.18%	61.10%
Test	70.20%	51.46%	59.39%

Table 5: Results of BERT locating SFs.

5.3 Performance of identifying ERs of a distractor using BERT.

In order to evaluate whether BERT model understands why a distractor is excluded, we modify BERT model and add a new component to perform the identification of distractors' ERs. The component is a multi-label (i.e., 8 labels including 1

ERs	#of options	R	P
Wrong details	484	1.65%	9.09%
Wrong temporal properties	29	13.79%	0.55%
Wrong subject-predicate-object triple	190	2.11%	5.88%
Wrong causality	218	24.77%	7.62%
Wrong necessary and sufficient conditions	106	1.89%	2.22%
Irrelevant to the question	85	17.65%	2.21%
Irrelevant to the article	352	4.83%	17.35%
Correct options	1984	28.18%	56.69%

Table 6: Results of BERT identifying ERs of distractors.

correct type and 7 types of ERs) classifier to predict the probability distribution of the types of ERs, and is jointly optimized with normal QA, and it shares the low-level representations. The classifier's objective is to minimize a cross entropy loss, which is jointly optimized with normal QA, and they share the low-level representations. For this task, the contextual input of BERT is ground-truth SFs, instead of the given article. The results are shown in Table 6. We observe that the performance of identification of ERs is quite low, indicating the significance and value of our dataset, which is to identify the limitations of existing systems on explainability and facilitates researchers to develop novel learning models.

5.4 Performance of different reasoning skills required by questions using BERT

We also investigate BERT's performance on the questions requiring different reasoning skills. We categorize the performance for each type of reasoning skills on the test set of GCRC. Table 7 shows the results. Note that the reasoning skills are annotated for options of questions. We observe BERT obtains the lowest score on the options requiring deductive reasoning. Overall, the system generally performs worse on each type of options, indicating the reasoning power of the system is not strong enough and needs to be significantly improved.

Reasoning	Detail understanding	Temporal/spatial	Coreference	Deductive
#of options	1683	42	179	143
Accuracy	31.97%	61.90%	39.66%	35.66%
	Mathematical	Cause-effect	Inductive	Appreciative
#of options	40	175	1056	130
Accuracy	60.00%	40.00%	33.14%	47.69%

Table 7: Results of BERT on GCRC by reasoning skills required for QA.

From the above, it can be seen that no baseline is realized to output answers, sentence-level SFs and ERs of a distractors together, but it doesn't affect our dataset to diagnose the limitations of existing RC models. For each task, we modify the existing model of BERT, output the corresponding explanations, and report their performance. Thus, the limitations of the model can be clearly identified. In the future, we will realize such a baseline that can do all the tasks together, and we will design a new joint metric for evaluating the whole question-answering process.

6 Conclusions

In this paper, we present a new challenging machine reading comprehension dataset (GCRC), collected from Gaokao Chinese, consisting of 8,719 high-level comprehensive multiple-choice questions. To the best of our knowledge, this is currently the most comprehensive, challenging, and high-quality dataset in MRC domain. In addition, we spend considerable effort to label three types of information, including sentence-level SFs, ERs of a distractor, and reasoning skills required for QA, aiming to comprehensively evaluate systems in an explainable way. Through experiments, we observe GCRC is very challenging data set for existing models, and we hope it can inspire innovative machine learning and reasoning approach to tackle the challenging problem and make MRC as an enabling technology for many real-world applications.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Key Research and Development Program of China (No.2018YFB1005103).

References

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shangmin Guo, Xiangrong Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2017. 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: Volume 1, Long Papers*, pages 111–120, Valencia, Spain. Association for Computational Linguistics.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zixian Huang, Yulin Shen, Xiao Li, Yu'ang Wei, Gong Cheng, Lin Zhou, Xinyu Dai, and Yuzhong Qu. 2019. GeoSQA: A benchmark for scenario-based question answering in the geography domain at high school level. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5866–5871, Hong Kong, China. Association for Computational Linguistics.

Naoya Inoue, Pontus Stenetorp, and Kentaro Inui. 2020. R4C: A benchmark for evaluating RC systems to get the right answer for the right reason. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6740–6750, Online. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219, Brussels, Belgium. Association for Computational Linguistics.

Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. 2017. Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 806–817, Vancouver, Canada. Association for Computational Linguistics.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. Investigating prior knowledge for challenging Chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155.

Shuohang Wang, Mo Yu, Jing Jiang, and Shiyu Chang. 2018. A co-matching model for multi-choice reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short*

Papers), pages 746–751, Melbourne, Australia. Association for Computational Linguistics.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach me to explain: A review of datasets for explainable NLP. *CoRR*, abs/2102.12060.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*.

Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5706–5713. AAAI Press.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Appendix

A. Hyper-parameters of Neural Baselines

The hyper-parameters of Co-Matching, BERT, XLNet are shown in Table 8, Table 9 and Table 10.

	GCRC	C3	RACE	DREAM
train batch size	8	16	16	32
dev batch size	8	16	8	32
Test batch size	8	16	8	32
epoch	50	100	50	100
learning rate	3e-5	1e-3	1e-3	2e-3
seed	128	64	256	128
dropoutP	0.2	0.2	0.2	0.2
emb dim	300	300	300	300
mem dim	150	150	150	150

Table 8: Hyper-parameters of Co-Matching

	GCRC	C3	RACE	DREAM
train batch size	32	16	16	32
dev batch size	4	16	4	16
Test batch size	4	16	4	16
len	320	384	450	384
epoch	6	3	10	3
learning rate	3e-5	1e-5	2e-5	1e-5
gradient accumulation steps	8	8	8	8
seed	42	42	42	42

Table 9: Hyper-parameters of BERT

	GCRC	C3	RACE	DREAM
train batch size	2	2	1	2
dev batch size	2	2	1	2
Test batch size	2	2	1	2
len	320	320	320	180
epoch	16	16	5	8
learning rate	2e-4	1e-3	2e-5	1e-5
gradient accumulation steps	2	2	24	2

Table 10: Hyper-parameters of XLNet