ReCO: A Large Scale Chinese Reading Comprehension Dataset on Opinion

Bingning Wang, Ting Yao, Qi Zhang, Jingfang Xu, Xiaochuan Wang
Sogou Inc.
Beijing, 100084, China
{wangbingning, yaoting, qizhang, xujingfang, wxc}@sogou-inc.com

Abstract

This paper presents the ReCO, a human-curated Chinese Reading Comprehension dataset on Opinion. The questions in ReCO are opinion based queries issued to commercial search engine. The passages are provided by the crowdworkers who extract the support snippet from the retrieved documents. Finally, an abstractive yes/no/uncertain answer was given by the crowdworkers. The release of ReCO consists of 300k questions that to our knowledge is the largest in Chinese reading comprehension. A prominent characteristic of ReCO is that in addition to the original context paragraph, we also provided the support evidence that could be directly used to answer the question. Quality analysis demonstrates the challenge of ReCO that it requires various types of reasoning skills such as causal inference, logical reasoning, etc. Current QA models that perform very well on many question answering problems, such as BERT (Devlin et al. 2018), only achieves 77% accuracy on this dataset, a large margin behind humans nearly 92% performance, indicating ReCO present a good challenge for machine reading comprehension. The codes, dataset and leaderboard will be freely available at https://github.com/benywon/ReCO.

Introduction

Machine reading comprehension (MRC), the ability to read the text and answer questions, has become one of the mainstreams in current natural language understanding (NLU) researches. Compared to other types of QA, MRC provided with only one document so the statistical information such as the number of answer occurrences could not be utilized, thus it requires a deeper understanding of the text. MRC has become an important part in many natural language processing applications, such as information retrieval (Nishida et al. 2018), event extraction (Ramamoorthy and Murugan 2018) and relation extraction (Levy et al. 2017).

One of the major contributions of the dramatic progress in MRC is the development of large scale corpus. Since the release of primal MCTest (Richardson, Burges, and Renshaw 2013), a great amount of datasets have been proposed, such as SQuAD (Rajpurkar et al. 2016; Rajpurkar, Jia, and Liang 2018), CNN/Daily Mail (Hermann et al. 2015), RACE (Lai et al. 2017), NarrativeQA (Kociský et al. 2018), etc. Based on these large scale datasets, a lot of deep learning based models have been built, such as BiDAF (Seo et al. 2017), QANet (Yu et al. 2018), etc. These models behave very well in MRC and some of them even surpass human performance.

However, despite the various types and relatively large scale, we found there are two main challenges previous MRC datasets has not been addressed:

1) The MRC context in most previous datasets are limit to the relatively long document or paragraph, which contains much irrelevant information to the question. Therefore, the *comprehension* process is sometimes reduced to the *retrieval* process (Sugawara et al. 2018), an MRC system could perform very well by merely finding the relevant sentences in a paragraph. For example, in SQuAD and NewsQA, the model's performance did not downgrade when only provided the sentence containing the ground truth answer (Weissenborn, Wiese, and Seiffe 2017; Min et al. 2018). In NarrativeQA (Kociský et al. 2018) or NaturalQuestions (Kwiatkowski et al. 2019) where the context passage is very long, the answer selection became the dominant factor for final result (Alberti, Lee, and Collins 2019; Kwiatkowski et al. 2019). The answer generation of MRC, which requires deep understanding of the text, has not been throughout evaluated.

2) Most previous MRC datasets are focus on factoid questions such as *who*, *when*, *where*, etc., so the candidate answers are limited to certain types such as *person*, *time*, *position*. Therefore, this kind of question does not require a complex understanding of language but merely recognizing the entity type could solve it properly. Sugawara et al. (2018) show that only using the first several tokens of the questions could achieve a significant improvement over random selection in many MRC datasets. This makes the reasoning process of machine learning methods built upon these datasets questionable (Jia and Liang 2017).

In this paper, we present ReCO, a large scale human-curated Chinese reading comprehension dataset focusing on opinion questions. In ReCO, the questions are real-world queries issued to search engine[1]. These queries are sampled and further filtered such that each query is a valid question

[1] https://www.sogou.com/

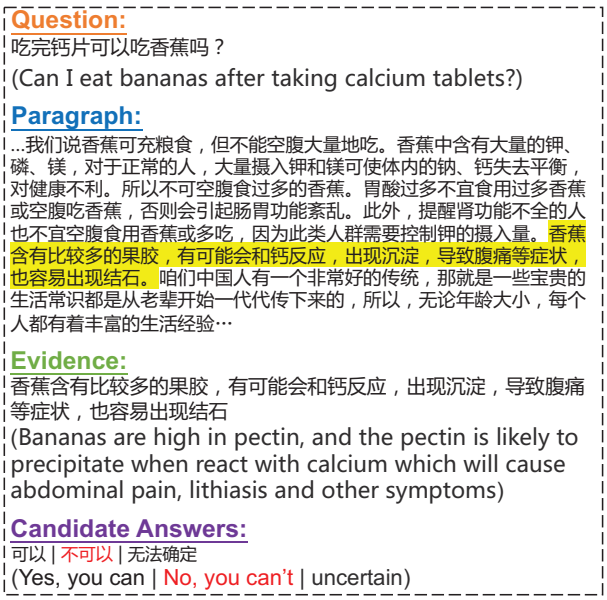

Figure 1: An example of ReCO. The evidence is extracted from the paragraph, which contains less irrelvant information to answer the question. The evidence may or may not consist of consecutive sentences in the paragraph.

and can be answered by yes/no/uncertain. Given the question, 10 retrieved documents are provided to the annotators, they were asked to select one document and extract the support evidence for that question. Finally, three candidate answers were given by the annotator: a positive one like *Yes*, a negative one like *No*, and an undefined one if the question could not be answered with the given documents. An example is shown in Figure 1.

Compared with previous MRC datasets, there are three main characteristics of the ReCO: First of all, in addition to the original document, we also provided the support evidence for the question. We do this because 1) Some documents may directly contain the answers with yes/no, where the answer could be trivially answered without understanding the subsequent evidence. 2) removing the irrelevant information could bypassing the answer selection error and concentrates the ReCO on the inference process of MRC. Data analysis shows that a large amount of ReCO questions require deep reasoning skills of text such as causal inference, logical reasoning, etc. (3), the $\langle$*paragraph, evidence*$\rangle$ could be utilized for further NLP applications such as summarization or answer selection.

Secondly, the questions in ReCO are opinion based real-world queries, which may be either factoid or non-factoid, and spanning many types of domains such as medical, health, etc. Besides, we use the search engine to obtain the passages which come from various resources, such as news articles, community QA or cyclopedia, etc. The diversity of questions and documents endows ReCO contains many aspects of world knowledge.

Finally, ReCO is very large and high-quality: it contains 300k questions, to our best knowledge it is the largest human-annotated opinion based QA dataset. In addition to the large scale, we introduce a rigorous inspection strategy to ensure the quality of the data, this makes ReCO relatively hard requiring deep understanding of text.

We applied several models on ReCO, including a modified version of BERT (Devlin et al. 2018) to fit the multiple-choice problem. The experimental results show that although it is a simple one-out-of-three problem, the best model only achieves 77% accuracy compared to humans 92%. The large gap between machine systems and humans indicates ReCO providing a good testbed for NLU systems.

Related Work

The MRC system from the NLP community could date back to 1990s when Hirschman et al. (1999) proposed a bag-of-words method that could give the answer to arbitrary text input. However, MCTest (Richardson, Burges, and Renshaw 2013) is widely recognized as the first dataset that we could build machine learning systems on it. Since the proposal of MCTest, there are more and more MRC datasets curated to facilitate MRC development. Table 1 shows an overview of these datasets and we divided them into three categories based on the answer type:

Multiple-Choices is the standard type of reading comprehension that contains several candidates. MCTest is a canonical multiple-choice dataset where each question is combined with 4 options. The MCTest is curated by experts and restricted to the concepts that a 7-year-old is expected to understand. Bioprocess (Berant et al. 2014) is another multiple-choice MRC dataset where the paragraph describing a biological process, and the goal is to answer questions that require an understanding of the relations between entities and events in the process. Other multiple choices MRC datasets including MCScript (Ostermann et al. 2018) that requires the system to understand the script of daily events, and RACE (Lai et al. 2017) where the questions are collected from the English exams of Chinese students.

Cloze is another type of MRC test in which some key points in the text are removed and should be filled given the contexts (Taylor 1953). Cloze could be deemed as complementary to multiple-choice reading comprehension for its reduced redundancy in text (Spolsky 1969). Hermann et al. (2015) use the article of CNN/Daily Mail as context, and blank out the entities in the summaries as the questions. Children's Book Test (CBT) (Hill et al. 2016) is another automatic generated cloze data. In CBT, a random entity was removed from a sentence and should be predicted given the previous 20 sentences. Clicr (Šuster and Daelemans 2018) is a medical domain cloze style data containing clinical case reports with about 100k gap-filling queries.

Open question answering is the dominant data type of current MRC where there are no options and the system must generate the answer. Most models in this types of dataset sometimes resort to the extractive strategy. SQuAD (Rajpurkar et al. 2016) is built upon Wikipedia where the context is a Wikipedia paragraph and the questions and answers were crowdsourced. SQuAD2.0 (Rajpurkar, Jia, and Liang 2018) is an extension to SQuAD that each docu-

dataset	type	question source	passage source	answer source	datasize	question type
SQuAD	OQA	human generated	wiki passage	human extracted	100k	F
SearchQA	OQA	J! Archive	search result	J! Archive	100k	F
MCTest	MC	human generated	story	human generated	2,000	NF
CNN/Daily Mail	CLOZE	abstract summary	news	entities	1.4m	F
CBT	CLOZE	children's book	children's book	entities	680k	NF
MARCO	OQA	query logs	search result	human generated	100k	F/NF
NarrativeQA	OQA	human generated	books and movie scripts	human generated	44k	NF
NaturalQuestions	OQA	queries logs	wiki document	human extracted	350k	F
DRCD	OQA	human generated	wiki document	human extracted	30k	F
DuReader	OQA	queries logs	search result	human extracted	200k	F/NF
ReCO	MC	queries logs	extraction from search result	human summarized	300k	F/NF

Table 1: Different MRC datasets. 'OQA', 'MC' refers to *open question answering* and *multiple-choice* respectively. *datasize* is the whole data size regardless the train/dev/test. F and NF denotes whether the question is factoid or non-factoid.

ment was given some questions that could not be answered. NewsQA (Trischler et al. 2017) is based on CNN/Daily Mail, the answers and questions are generated by different people to solicit exploratory questions that require reasoning. NaturalQuestions (Kwiatkowski et al. 2019) is a Wikipedia based dataset focusing on factoid questions. In SearchQA (Dunn et al. 2017) and MARCO (Nguyen et al. 2016) the documents were collected from search engine.

ReCO is also related to another Chinese MRC dataset such as DRCD (Shao et al. 2018), CMRC (Cui et al. 2018) and DuReader (He et al. 2018). Specifically, DuReader also contains yes/no questions. However, it only contains a small portion (8%) of the yes/no questions, and only the whole documents are provided as context, which contains much more irrelevant information, or may directly answer the questions without deep reasoning of the evidence.

Compared with other datasets, ReCO is an opinion based MRC dataset focusing on the yes/no/uncertain questions. The questions and context are obtained from real-world queries and web pages which shows diversity in domains. Besides, the context passage in ReCO is very short evidence and in most cases, deep inference skills such as analogue, logical reasoning are required to answer the question. ReCO is also related to recognizing textual entailment (**RTE**) where the task is to determine whether there exits entailment/neutral/contradiction relation between two sentences, such as SNLI (Bowman et al. 2015) and MNLI (Williams, Nangia, and Bowman 2017). However, entailment is a more narrow concept that the truth of hypothesis strictly requires the truth of the premise, whereas in ReCO the *premise* is the evidence and *hypothesis* is the question, so the inference between them contains much broader concepts such as deduction, induction, and abduction.

Data Collection

The data collection process includes a query selection, passage retrieval, passage filtering, evidence extraction, and answer generation process. Rigorous inspections are applied to ensure the quality and difficulty of the datasets. The conceptual scheme is illustrated in Figure 2.

Qustion Curation

Intent Analysis: First of all, we sample 10 million queries issued to Sogou Search Engines. Next, we use the off-the-shelf intent analysis system to determine whether the query is a valid question. Then we drop some queries that contain sex, violence and other inappropriate content. These two processes exclude nearly 95 percent queries.

Query Filtering: Given the filtered queries, we build a simple symbolic feature based machine learning system to determine whether the question could be answered by yes or no. The features we use are whether it contains 'Could I', 'whether', 'Is there any', etc. This simple system is very effective that only a small fraction of questions do not present the yes/no query intent.

After the intent analysis and query filtering process, we obtain the original questions, although some invalid questions may pass through the above filtering processes. The next several steps could further reject some of these questions to make invalid questions as less as possible.

Document Collection

Document Retrieval: we use the off-the-shelf Sogou search engine to retrieve 10 pages for each question, and then extract the body content of each page. The main focus of ReCO is text understanding but not fact seeking, so we did not filter out the pages from the forums or community sites where the answers may be subjective.

Document Filtering: is proposed to prevent the retrieved documents containing some trivial answers that perfectly match the question. For example, if the question is 'Can pregnant women eat celery?', it would be meaningless to give the candidate document which contains 'pregnant women can eat celery'. We use some word-based rules to remove the documents that contain significant surface overlap with the question.

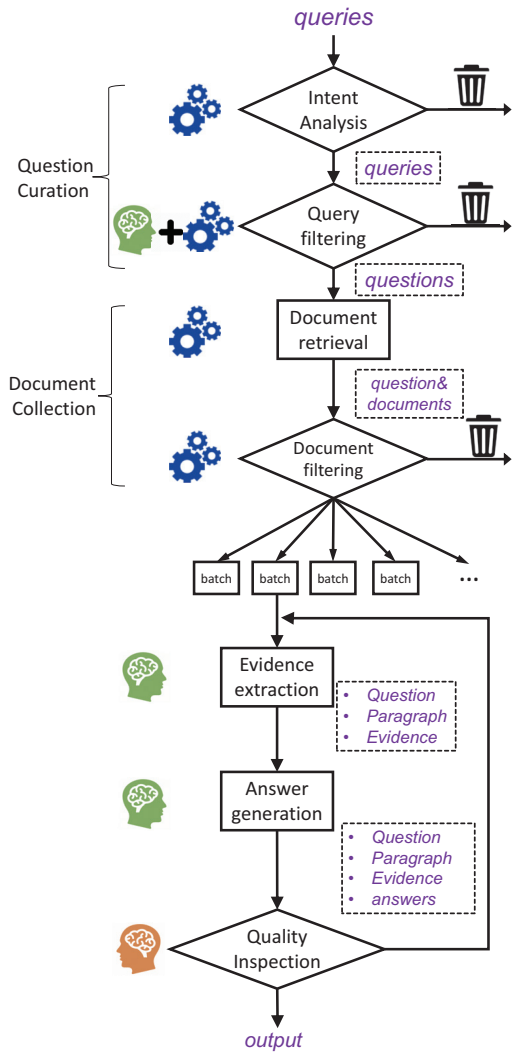

Figure 2: The data collection process of ReCO. refers the annotator, refers the authoritative checker, refers to off-the-shelf system or machine learning models.

Evidence & Answer Curation

We first randomly divided the question-documents pairs into a lot of batches with each batch contains 5k samples. Then each batch is annotated by a single annotator by the following processes.

Evidence Extraction: given a question and its relevant documents, we ask the annotator to extract the snippet from the document as the evidence. And the document containing the evidence is the context paragraph. There are four principles for this extraction:

- The evidence should be self-contained to answer the question and as short as possible.
- If multiple support evidence could answer the question, select the most implicit one requiring deeper inference.
- If there is no evidence in the document that could answer the question, select the most relevant passage.
- If a question could not be answered by yes/no/undefined, it should be rejected.

The first principle is introduced so that the extracted evidence should contain less irrelevant information to the question, and therefore bypass the answer selection errors which is the bottleneck in some other datasets. The second principle ensures the difficulty of the extracted passage, enabling deeper reasoning of the text.

Answer Generation: The annotator was asked to give abstractive candidates to the answer after the evidence extraction process. It should contain a positive one such as 'Can' and a negative one such as 'Cannot', and an undefined one if the question could not be answered by the passage[2]. The answer candidates are summarized by the annotator that it may not be the words in the original evidence or question which are shown in Figure 3.

(a) negative answers (b) positive answers

Figure 3: Wordcloud of candidate answers. The common positive answers are *can, need, yes, useful*. And the most common negative answers are *can't, unable, not, wouldn't,*.

Quality Inspection

After the above processes, we obtain a lot of batches, each batch is examined by the expert checkers who are expert at our domain and fully understand the demanding quality of the dataset. There are four key rules for the expert checker to determine whether a sample is false:

- ⋆ The answer is incorrect.
- ⋆ The question is blurred, or it could not be answered by yes/no/undefine.
- ○ The evidence has much irrelevant information.
- ○ The question is too easy given the evidence.

[2] as we may remove the valid documents in the document filtering process, or the question is too unusual to get a good answer. In this case, the passage is selected as the most relevant snippet in the documents to the question.

The rules with ⋆ are the strict rules that the instance is absolute false. The rules with ∘ are the loose rules that this instance is only *half wrong* and only account 0.5 in the final error summation.

For each batch, we randomly sample a fraction of instances and send it to the expert checker. The checker examines the quality of these data based on the above rules. If the accuracy of a batch is higher than 0.95, it is passed and accepted to the final set. Otherwise, this batch is rejected and pushed back to the evidence extraction process of the corresponding annotator and relabeled again.

Test Data Collection

The test data requires higher quality compared with the training set for its evaluation usage. After the document filtering process, we sent each sample to two annotators and annotated independently. If the answer is the same across two annotators, it was sent to the third annotator to select the evidence provided by the preceding two annotators.

Finally, we obtain 280,000 training data and 20,000 testing data. The average length of paragraph, evidence and question is 924.5 characters, 87.1 characters and 10.6 characters respectively. The ratio of positive, negative, undefined questions is about 5:4:1.

Dataset Analysis

To understand the properties of ReCO, we sample 200 instances to analyze three aspects of ReCO: (i) The diversity of the question domains. (ii) The domains of the evidence. (iii) The reasoning skills required to answer the questions.

Question Domains

The diversity of question domains could somewhat reflect the world knowledge coverage of the data. We divided the questions into 5 categories: (1) Health: about disease, foods, exercise, etc. (2) SciTech: including science, technique, tools, etc. (3) Society: questions about legal provisions, stipulations, education, etc. (4) Life: questions about life such as public transport information, vacation, shopping, etc. (5) Culture: about literature, art, history, etc. In Table 2, we can see that the question domain is varied. This is an advantage over some previous works such as SearchQA, or NarrativeQA that the question is focused on specific domain. The diversity in ReCO question makes it a comprehensive dataset containing many aspects of the world knowledge.

Evidence Domains

The documents in ReCO are derived from the search engine, so pages from every possible domain could be the candidate answer. We further analyze the diversity of the evidence domains. We coarsely divide the evidence into three categories based on its source document: (1) Vertical: sites providing authoritative information for specific domain, such as medical center, government legislation department, etc. (2) Forum: the online discussion site where people can hold conversations in the form of posted messages. (3) Others: including the evidence from other sources such as encyclopedia, news articles, etc.

Domain	Example	Percentage(%)
Health	晒太阳真的能补钙吗? (Could sunbathing improve Calcium supplementation?)	44.5
SciTech	u盘进水还能用吗? (Can usb sticker work after falling into water)	12.5
Society	湖南高考分数线高吗? (Is the score of Hunan college entrance examination high?)	17.5
Life	移动营业厅星期天开门吗? (Does the mobile business hall open on Sunday?)	20.5
Culture	西方人吃米饭吗 (Do westerners eat rice?)	5.0

Table 2: Question domains of ReCO.

Domain	Example	Percentage(%)
Vertical	FDA, MamaBang	40.0
Forum	Quora, Sogou Wenwen	44.5
Others	Sogou Baike, People's daily	15.5

Table 3: Evidence Domains of ReCO. The classification is based on the source document of the evidence.

In Table 3, we can see the evidence domain of ReCO is diverse, including both formal texts such as vertical sites articles and informal text such as forum discussions. This is an advantage over some other datasets of which the domains are limited to certain types, such as Wikipedia (Rajpurkar et al. 2016; Kwiatkowski et al. 2019), news articles (Hermann et al. 2015) or children stories (Hill et al. 2016), which elude understanding the stylistic feature of text and hard to test the generalization ability of existing models. The diversity in passage domains is the prerequisite of literary form understanding which is a key point in reading comprehension (Snow 2002; Makhoul and Copti-Mshael 2015).

Reasoning Skills

To achieve natural language understanding, reasoning skills are the main desiderata of MRC datasets compared with other factors. To get a better understanding of reasoning skills required to answer the questions in ReCO, following previous works on reading comprehension (McNamara and Magliano 2009) and MRC (Sugawara et al. 2017; 2018), we identified 7 reasoning skills and classify them into shallow (♢) or deep(♣)[3]:

1) Lexical Knowledge♢: lexical information about the word, such as synonymy, hypernymy or morphology.

2) Syntactic Knowledge♢: is the knowledge about sentence structure, such as part of speech, apposition or clause relation such as coordination or subordination, including relative clauses.

3) Coreference♣: is a skill to track some objects, including anaphora and cataphora.

4) Casual Inference♣: is the knowledge about causality between the cause and effect, which are sometimes repre-

[3]Note that this categorization is empirical and provisional without very solid theoretical background.

Reasoning Type	Example	SQuAD	MARCO	NewsQA	DuReader	ReCO
1) Lexical Knowledge$^{\diamond}$	Q:榴莲籽可以吃吗 P:榴莲的种子富含蛋白质 Q: Could I eat the seed of durian P: The germ of durian is high in protein.	70.0	63.0	84.0	32.5	26.5
2) Syntactic Knowledge$^{\diamond}$	Q:脑梗可以吃牛肉吗 P:牛肉在得了脑梗后可以吃 Q:Could I eat beef after stroke? P:Beef is edible after stroking.				31.0	32.0
3) Coreference$^{\clubsuit}$	Q:白色和黄色是一样的味道么? P:...白色是淡香精，黄色是香精，它们的味道... Q:the yellow one and white one have same smell? white ..., yellow, their smell is ...	13.0	15.0	**24.0**	21.5	10.0
4) Casual Inference$^{\clubsuit}$	Q:环球黑卡能不能刷卡? P:环球黑卡是标识卡，标识卡不能刷卡... Q:Could I pay by Global black card?P:Global black card is ID card, ID card couldn't be used for paying...	0.0	0.0	4.0	17.0	**35.5**
5) Ellipsis$^{\clubsuit}$	Q:上海的网吧可以抽烟么? P:抽烟是命令禁止的，网吧内一般不可以 Q:Could I smoking in the Internet bar of Shanghai? P:Smoking is forbidden in anywhere, Internet bar is included.	3.0	2.0	15.0	21.0	**29.0**
6) Logical Reasoning$^{\clubsuit}$	Q:所有的茶都是碱性的吗?P:乌龙茶、红茶呈弱酸性 Q:Are all the teas alkaline? P:Oolong and black tea is acidic.	0.0	1.0	2.0	4.0	**18.5**
7) Specific Knowledge$^{\clubsuit}$	Q:石家庄属于北方吗? P:秦岭淮河以北的就是北方 Q: Is Shijiazhuang located in North China?: P:North of the Huaihe River is the northern China.	26.0	14	**29.0**	24.0	21.5

Table 4: Frequencies (%) of the shallow ($^{\diamond}$) or deep($^{\clubsuit}$) reasoning skills required for MRC datasets. The results of SQuAD, MARCO and NewsQA are borrowed from Sugawara et al. (2017). We exclude the 'undefined' questions because the evidence may not relate to the question. Note that some question requires more than one skill.

sented by events.

5) Ellipsis$^{\clubsuit}$: recognizing omitted information (argument, predicate, quantifier, place, etc).

6) Logical Reasoning$^{\clubsuit}$: understanding the predicate logic such as negation, conditionals, quantifiers, transitivity.

7) Specific Knowledge$^{\clubsuit}$: is the skills in specific domain, temporal, mathematical, spatiotemporal among others.

The statistics of different reasoning skills are shown in Table 4. We can see that compared with other prevalent MRC datasets, a large amount of questions in ReCO requires deeper reasoning skills. Specifically, we found that many questions require casual inference and ellipsis which are very difficult for current systems. Reviewing the data collection process, we conclude two factors making the ReCO more difficult:

Intrinsic: In ReCO, we confine the answer type to be yes/or/uncertain, thus the question answering can be cast as an entailment recognition process. However, as the questions are real-world queries generated by users and the evidence is the extraction of the document retrieved by the search engine, there is no direct correspondence between the '*premise*', i.e. the evidence, and the '*hypothesis*'-the question. This is in contrast with many previous MRC datasets, where the questions are generated based the context documents, so there is a strong correlation between the question and context in previous datasets that sometimes trivial to answer. The information decoupling between the evidence and question in ReCO necessitate the deeper understanding of the textual inference.

Extrinsic: the data collection process involves a rigorous quality inspection step that some data would be eliminated if it was too easy. Although this step may result in a large decrease of the data size, we believe it is indispensable to remove too easy samples given that the MRC systems are prone to be attacked by adversarial examples (Jia and Liang 2017), for which the simple pattern in the data is the main reason (Sugawara et al. 2018).

Experiments

Baselines

To evaluate the baseline performance, we consider three competitive models that perform very well in MRC and many other NLU applications:

- **BiDAF** (Seo et al. 2017) is the very first deep learning model achieving remarkable performance on MRC. It is built upon LSTM and bi-directional attention was introduced to interact the question and the answer.
- **BiDAF***: In addition to BiDAF, ELMO (Peters et al. 2018) is introduced for word representation initialization. ELMO is an LSTM based bi-directional language model trained on unsupervised data which shows advantage compared to the word embedding methods.
- **BERT** (Devlin et al. 2018) is a recently proposed model that substantially advanced the state-of-the-art in many NLP tasks. It is based on Transformer (Vaswani et al. 2017) and pre-trained in large unlabeled data, the objective contains a mask language modeling task and a next sentence prediction task.

The original BERT model is not designed for the multiple-choice problem, so we modified the architecture to take the candidate answers into account. Concretely, we concatenate the candidate answers after the sequence with a special `[CLS]` token. Then we use the BERT to represent the whole sequence. Finally, the output representations of the three `[CLS]` tokens are used to predict the three candidate answers probability. The loss is the cross entropy.

Metric	SQuAD F1	SNLI ACC	MARCO Rouge-L	DuReader Rouge-L	ReCO-paragraph ACC	ReCO ACC
Random	0*	33.3	0*	0*	33.2	33.2
BiDAF	77.2	86.7	23.9	39.0	55.8	68.3
BiDAF*	81.1	88.7	43.6	52.9	58.9	70.9
BERT_b	88.5	90.2	48.2	53.4	61.1	73.4
BERT_l	91.8	90.8	-	-	65.3	**77.0**
Human	91.2	87.7*	53.9	57.4	88.0	91.5

Table 5: Result of different models on ReCO and other MRC datasets. * means the estimated results. ReCO-para denotes we use the original paragraph as the context. BERT$_{_b|_l}$ denotes BERT base or large model. The results of SQuAD, SNLI, DuReader and MARCO were derived from their leaderbord or paper.

Common Setup

We use the ReCO training set to build sentencepiece (Kudo and Richardson 2018) tokenizer and set the vocabulary size to 35,000. For BiDAF, we adopt the same experimental setting with the original implementation. For BERT and ELMO, we use the openly released code[4]. In all experiments we set the batch size to 48 and run on 8 Nvidia V100 GPUs.

Table 5 shows the result of these models on ReCO. Compared with their performance on other MRC datasets, It is clear that the current best models still struggle to achieve a good result in ReCO, even though it is a simple three-category classification problem and a random system could achieve 1/3. Needless to say the input is the short noise-free evidence. If we fed the long document instead of the evidence to the model, the results drop a lot, which means the answer selection process also plays a key role in MRC.

We analysis the error of BERT large model on ReCO based on the inference skills in Table 4. The result is shown in Figure 4. We can see that as Bert has been pre-trained with large unlabeled data, the lexical and syntactical knowledge has been well modeled and the corresponding accuracy is high. But the deeper inference skills, such as logical reasoning or specific knowledge, is not directly present in data so the performance of these questions is not satisfied. Incorporating more sophisticated knowledge, such as word sense (Levine et al. 2019) or knowledge base information may further benefit the model, which we leave for future study.

RI index

To understand the difficulty of a dataset, we proposed a relative improvement (**RI**) index:

$$\mathrm{RI} = \frac{S_{\mathrm{model}} - S_{\mathrm{random}}}{S_{\mathrm{human}} - S_{\mathrm{random}}} \quad (1)$$

S_{model}, S_{random} and S_{human} denote the score of best machine learning model, the score of a random system and the score of human beings, respectively. RI is measured by how much improvement the best machine learning models have achieved compared to how much improvement the human have achieved w.r.t. a random system. This index reflects the gap between the performance of current systems and human beings, and thus a criterion indicating the difficulty.

[4]https://github.com/pytorch/hub

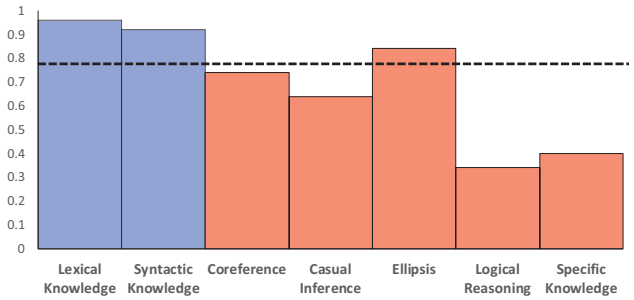

Figure 4: The accuracy of BERT large model on different types of questions based on reasoning skills in Table 4. The dotted line is the average accuracy.

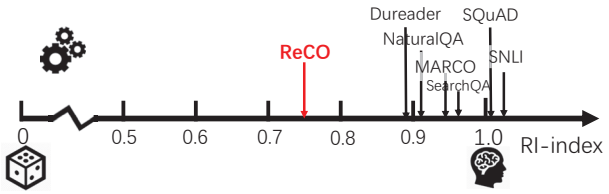

Figure 5: RI-index of different datasets. Results of other datasets obtained from their paper or leaderboard.

Figure 5 shows the RI indexes of different MRC datasets. It is clear that the machine learning models have achieved competitive results in most datasets, some of them even surpass human performance. But for ReCO, there is still a large headroom for machines to improve. On the one hand, it reflects that the ReCO is a relatively hard task that the current model is still incompetent. On the other hand, most ReCO questions require deep reasoning skills, so new mechanisms should be introduced to the MRC models to achieve higher level inference, such as logical reasoning, etc.

Conclusion

This paper presents ReCO, a large scale opinion based Chinese reading comprehension dataset contains 300k questions. We use a very rigorous data inspection process to guarantee the quality of the data. ReCO contains short evidence which bypasses the answer selection error, and data analysis shows that most of the questions require deep reasoning skills. We develop a relative improvement index to measure the difficulty of the dataset. Experimental results and RI index demonstrate the difficulty of ReCO. Much more efforts should be made to filling the gap between machines and humans in this text understanding application.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. We thank the Sinovation Ventures for hosting the AI Challenger 2018 competetions which introduce this dataset to the chinese NLP community.

References

Alberti, C.; Lee, K.; and Collins, M. 2019. A bert baseline for the natural questions. *arXiv preprint arXiv:1901.08634*.

Berant, J.; Srikumar, V.; Chen, P.-C.; Vander Linden, A.; Harding, B.; Huang, B.; Clark, P.; and Manning, C. D. 2014. Modeling biological processes for reading comprehension. In *EMNLP*.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *EMNLP 2015*.

Cui, Y.; Liu, T.; Chen, Z.; Ma, W.; Wang, S.; and Hu, G. 2018. Dataset for the first evaluation on chinese machine reading comprehension. In *LREC 2018*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dunn, M.; Sagun, L.; Higgins, M.; Guney, V. U.; Cirik, V.; and Cho, K. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

He, W.; Liu, K.; Lyu, Y.; Zhao, S.; Xiao, X.; Liu, Y.; Wang, Y.; Wu, H.; She, Q.; Liu, X.; Wu, T.; and Wang, H. 2018. Dureader: a chinese machine reading comprehension dataset from real-world applications. In *QA@ACL*.

Hermann, K. M.; Kociský, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. In *NIPS*.

Hill, F.; Bordes, A.; Chopra, S.; and Weston, J. 2016. The goldilocks principle: Reading children's books with explicit memory representations. *CoRR* abs/1511.02301.

Hirschman, L.; Light, M.; Breck, E.; and Burger, J. D. 1999. Deep read: A reading comprehension system. In *ACL*, 325–332. Association for Computational Linguistics.

Jia, R., and Liang, P. 2017. Adversarial examples for evaluating reading comprehension systems. In *EMNLP*.

Kociský, T.; Schwarz, J.; Blunsom, P.; Dyer, C.; Hermann, K. M.; Melis, G.; and Grefenstette, E. 2018. The narrativeqa reading comprehension challenge. *TACL* 06:317–328.

Kudo, T., and Richardson, J. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Kwiatkowski, T.; Palomaki, J.; Rhinehart, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Kelcey, M.; Devlin, J.; et al. 2019. Natural questions: a benchmark for question answering research.

Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; and Hovy, E. H. 2017. Race: Large-scale reading comprehension dataset from examinations. In *EMNLP*.

Levine, Y.; Lenz, B.; Dagan, O.; Padnos, D.; Sharir, O.; Shalev-Shwartz, S.; Shashua, A.; and Shoham, Y. 2019. Sensebert: Driving some sense into bert. *CoRR*.

Levy, O.; Seo, M.; Choi, E.; and Zettlemoyer, L. S. 2017. Zero-shot relation extraction via reading comprehension. In *CoNLL*.

Makhoul, B., and Copti-Mshael, T. 2015. Reading comprehension as a function of text genre and presentation environment: comprehension of narrative and informational texts in a computer-assisted environment vs. print. *Psychology* 6(08):1001.

McNamara, D. S., and Magliano, J. 2009. Toward a comprehensive model of comprehension. *Psychology of learning and motivation* 51:297–384.

Min, S.; Zhong, V.; Socher, R.; and Xiong, C. 2018. Efficient and robust question answering from minimal context over documents. *arXiv preprint arXiv:1805.08092*.

Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; and Deng, L. 2016. Ms marco: A human generated machine reading comprehension dataset. *NIPS*.

Nishida, K.; Saito, I.; Otsuka, A.; Asano, H.; and Tomita, J. 2018. Retrieve-and-read: Multi-task learning of information retrieval and reading comprehension. In *CIKM*.

Ostermann, S.; Modi, A.; Roth, M. A.; Thater, S.; and Pinkal, M. 2018. Mcscript: A novel dataset for assessing machine comprehension using script knowledge. *CoRR* abs/1803.05223.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. S. 2018. Deep contextualized word representations. In *NAACL-HLT*.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know what you don't know: Unanswerable questions for squad. In *ACL*.

Ramamoorthy, S., and Murugan, S. 2018. An attentive sequence model for adverse drug event extraction from biomedical text. *CoRR*.

Richardson, M.; Burges, C. J.; and Renshaw, E. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *EMNLP*.

Seo, M. J.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2017. Bidirectional attention flow for machine comprehension. *ICLR*.

Shao, C.-C.; Liu, T.; Lai, Y.; Tseng, Y.; and Tsai, S. 2018. Drcd: a chinese machine reading comprehension dataset. *ArXiv*.

Snow, C. 2002. *Reading for understanding: Toward an R&D program in reading comprehension*. Rand Corporation.

Spolsky, B. 1969. Reduced redundancy as a language testing tool.

Sugawara, S.; Kido, Y.; Yokono, H.; and Aizawa, A. 2017. Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In *ACL*, volume 1, 806–817.

Sugawara, S.; Inui, K.; Sekine, S.; and Aizawa, A. 2018. What makes reading comprehension questions easier? *EMNLP*.

Šuster, S., and Daelemans, W. 2018. Clicr: a dataset of clinical case reports for machine reading comprehension. *NAACL*.

Taylor, W. L. 1953. "cloze procedure": A new tool for measuring readability. *Journalism Bulletin* 30(4):415–433.

Trischler, A.; Wang, T.; Yuan, X.; Harris, J.; Sordoni, A.; Bachman, P.; and Suleman, K. 2017. Newsqa: A machine comprehension dataset. In *Rep4NLP@ACL*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*.

Weissenborn, D.; Wiese, G.; and Seiffe, L. 2017. Making neural qa as simple as possible but not simpler. In *CoNLL*.

Williams, A.; Nangia, N.; and Bowman, S. R. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Yu, A. W.; Dohan, D.; Luong, M.-T.; Zhao, R.; Chen, K.; Norouzi, M.; and Le, Q. V. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *ICLR*.