Abstract
Recent studies raised that reading comprehension (RC) models learn to exploit biases and annotation artifacts in current Machine Reading Comprehension (MRC) datasets to achieve impressive performance. This hinders the community from measuring sophisticated understanding of RC systems. MRC questions whose answers can be rightly predicted without understanding their contexts are defined as biased ones. Previous researches aimed to split unintended biases and determine their influence have some limitations. Some methods using partial test data to extract biases lack holistic consideration with question-context-option tuple. Others relied on artificial statistical features are limited by question types. In this paper, we employ two simple heuristics to identify biased questions in current MRC datasets through human-annotated keywords. We implement three neural networks on the biased data and find that they have outstanding abilities to capture the biases, and further study the superficial features of the biased data exploited by models as shortcuts in views of lexical choice and paragraphs. Experiments show that (i) models can answer some questions merely using several keywords which are unanswerable or difficulty for human. (ii) lexical choice preference in options creates biases utilized by models. (iii) fewer paragraphs are more likely to introduce biases in MRC datasets.

1 Introduction
Machine Reading Comprehension (MRC) as a critical task in many real-world applications requires machines to answer a question by understanding the given context (Hirschman et al., 1999). Numerous MRC datasets have been published and facilitated the progress of MRC models. Although recent state-of-the-art models have reached impressive performance, it does not indicate they have possessed human-like reading comprehension capabilities (Jia and Liang, 2017). Data collection is the most under-scrutinized step of the machine learning pipeline (Paritosh, 2020). Moreover, human-annotated datasets usually contain biases exploited by neural networks as shortcut solutions to achieve high accuracy (Schwartz et al., 2017).

Previous study (Yu et al., 2020) fed models with only option data and treated the correctly predicted ones as biased while lacking attention to the contexts. Sugawara et al. (2018) extracted biased data through artificial features restricted by question expressions. We conjecture that biases exist in not only options but questions and articles and concern that what features resulting in such biases and acting as shortcuts for models. To this end, this article aims to investigate biases exist in current MRC datasets and summarize suggestions for future MRC dataset. We define MRC questions whose answers can be rightly predicted without understanding their contexts as biased ones.

The contributions of this paper are as follows. Firstly, we introduce a Human-Inspired Chinese Reading Comprehension (HICMRC) dataset with high-quality complex reasoning multi-choice questions from Chinese standard examinations, and collect human results and manually labelled token-level supporting facts related to questions in passage for explainable evaluation. Secondly, we evaluate three baseline models and extract biased datasets through two filtering heuristics. Finally, we analyze superficial features in the biased datasets by comparison with non-biased ones and summarize recommendations for future MRC data construction.

2 Related Work
Levesque (2014) proposed that we should avoid building problems that can be solved by matching patterns, using unintended biases, and choice constraints when testing AI. Min et al. (2018) observed that 92% of answerable questions in SQuAD can be predicted merely using a single context sentence.
Agrawal et al. (2016) studied the behavior of models by variable length of the first question tokens in the field of visual QA. Wiegreffe and Marasovic (2021) concluded three prevalent neural networks divided explanations into two categories as justification and introspective. For MRC tasks, MultiRC (Khashabi et al., 2018) and HotpotQA (Yang et al., 2018) provided sentence-level SFs regarded as justifications and introspective explanations. Inoue et al. (2020) divided explanations into two categories as justification and retrospective. For MRC tasks, MultiRC (Khashabi et al., 2018) and HotpotQA (Yang et al., 2018) provided sentence-level SFs regarded as justification explanations. R4C (Inoue et al., 2020) and 2WikiMultiHopQA (Ho et al., 2020) offered both justification and introspective explanations. There exist fewer Chinese datasets with explanation information and most of them were collected from standard Chinese exams. C3 (Sun et al., 2020) questions were provided with types of essential prior knowledge. GCRC (Tan et al., 2021) labelled three kinds of information including supporting facts, error reasons and types of reasoning skills. Inspired by these datasets, we spent tremendous effort to design a credible annotation method and collected token-level supporting facts relevant to questions in context for explainable model evaluation and biased data analysis.

### 3 Data Collection and Baselines

#### 3.1 Data from Examined Datasets

HICMRC’s data format is similar to other multiple-choice RC datasets like Sun et al. (2020), where each instance consists of a context, a question, three distracters and a right option. We have spent tremendous effort to construct challenging high-quality questions for testing advanced passage-level MRC abilities. Firstly, we filtered samples from C3M test set by a series of rules (see details in Appendix A). Secondly, C3 has shorter document and easier questions since it is collected from Chinese-as-a-second-language exams, we replenished samples from Chinese Junior Middle School Modern Reading Exams following the preceding rules. Then we invited experts to proofread passages, rectify mistakes like typos, and examine the questions cannot be easily guessed by comparisons among options or without understanding context. Finally, we adjusted answers’ labels so that they are evenly distributed in A/B/C/D and summarized the statistics of HICMRC test dataset (200 documents and 200 questions in total) in Table 1.

#### 3.2 Human Results and Annotations

We obtained human performance by inviting 48 undergraduates to complete 60 questions in HICMRC, where they were asked to read a question first, then its corresponding passage and answer it among the shown options. For more comprehensive analysis on biased data and explainable evaluation of models, we also hired 66 undergraduates to annotate token-level supporting facts in passages which are crucial for answering their corresponding questions. We would emphasize that the annotation task is extremely challenging since annotations are evaluated by plausibility (how well annotations support prediction) and faithfulness (how accurately annotations represent the decision process) (Yang et al., 2019). Consequently, we took enormous effort to design the annotation procedure and attach them in Appendix B.

#### 3.3 Baseline Systems

We implemented three prevalent neural networks to get models’ performance including fastText, Co-matching and Chinese Bert-Base, which have reached promising results on MRC task according to previous researches (Joulin et al., 2017; Wang et al., 2018; Li et al., 2018; Devlin et al., 2018).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Human Answerable set</th>
<th>Human Unanswerable set</th>
</tr>
</thead>
<tbody>
<tr>
<td>FastText</td>
<td>0.445</td>
<td>0.395</td>
</tr>
<tr>
<td>Co-matching</td>
<td>0.26</td>
<td>0.54</td>
</tr>
<tr>
<td>BERT</td>
<td>0.433</td>
<td>0.66</td>
</tr>
<tr>
<td>Human</td>
<td>0.353</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Table 1: Statistics and reading comprehension accuracy of models and human on four datasets.
We first train three models using C3M training data with consistent parameters as in C3. For evaluation, we run every experiment five times and report models with the best development set performance. Details of the baselines and implementation are in the Appendix C.

Table 1 shows comparison results. We observe that both human and models underperform on HICMRC test data than C3M test which suggests that HICMRC is more challenging. Additionally, human performs worse when using keywords rather than complete passage as inputs (0.78 to 0.72 in accuracy) while Bert’s accuracy increases from 0.493 to 0.532. Co-matching and fastText were slightly affected with drops of 0.03 and 0.035. The inconsistent trends between human and models indicating that there may exist biases learned by models. Meanwhile, we split answerable and unanswerable subsets by human accuracy and it is interesting that the performance gap of models between two subsets disagrees with that of human.

4 Experiments

4.1 Filtering Heuristics to find biased data

Recent studies have exposed that datasets created by experts may introduce biases and models can utilize the biases to achieve high accuracy without truly understanding the context (Yu et al., 2020). One goal of this paper is to identify the biases in HI-CRM for more comprehensive model evaluation. We filtered out biased data based on the influence of two filter heuristics: (i) Human-performance-based. (ii) Context-aware, and then investigated baseline models’ performance on biased and non-biased subsets. Several biased examples are given in Appendix D.

Human-performance-based Heuristic. As shown above, models perform relatively inconsistent or even reverse on human answerable and unanswerable subsets compared with human. Some previous work identified questions that can be rightly predicted when removing the context and question in the inputs (Yu et al., 2020), which neglected biases in passages and questions. To this end, we feed masked passage, its corresponding question and options into three baseline models for each data point. In this way, we identify questions that are Unanswerable for Human (UH) while can be correctly Answered by Models merely using annotated Keywords (AMK) and other consistent inputs. We believe that such data exists unintended biases or shortcuts exploited by models but neglecting by human, and donate them as $D_{biased}^{1} = UH \cap AMK$.

Context-aware Heuristic. This heuristic is to detect questions that are Unanswerable for Models after reading complete Context(UMC) but Answered by merely reading annotated Keywords(AMK). In other words, questions that are answerable by hints from human annotations cannot examine model abilities of understanding of the context and locating relevant information for answering questions, which donated as $D_{biased}^{2} = UMC \cap AMK$.

To investigate what makes MRC questions fail to test models’ sophisticated MRC abilities to answer beyond using superficial cues, we examine the following statistical characteristics on biased and non-biased data. Biased data is formulated as $D_{biased} = D_{biased}^{1} \cup D_{biased}^{2}$. For more precise comparative analysis, we remove questions that can be correctly answered both by human and models either using keywords or full context. Namely, non-biased data contains Unanswerable questions for Models neither with complete Context (UMC) nor annotated Keywords (UMK), which is expressed as $D_{non-biased} = UMC \cap UMK$.
4.2 Experiments between biased and unbiased data

4.2.1 Lexical Choice in Options
Following method in an English counterpart Dataset RECLOR (Yu et al., 2020), we investigate the biases of lexical choice in options. For the character-level tokens in options, we compute their conditional probability of label $l \in \{\text{right, wrong}\}$ given token $t$, where $p(l|t) = \text{count}(l,t)/\text{count}(t)$. The larger $p$ value for a token, the greater its contribution to the prediction of corresponding options (Poliak et al., 2018). Table 2 presents character-level tokens with the largest $p$ scores which occur at least twenty times (considering many tokens with largest $p$ values are of low frequency) in biased and unbiased data based on the performance of human and baseline model Bert. We notice that lexical choice of right options in biased data obviously differs from the of data and is more concentrated to some particular tokens with higher $p$ scores.

4.2.2 Token-level Supporting Facts Distribution
To explore biases resulting in $D_{biased}^2$, where questions are unanswerable with original passage-question-option tuple but can be correctly predicted using annotated keywords, we focus on the analysis of annotated keywords distribution in passages. We separately count the number of different sentences and paragraphs in which keywords are distributed for each passage and perform a significance test to determine whether sentences/paragraphs position distribution of keywords contributes to performance gap of models. Table 3 represents the average/maximal number of sentences and paragraphs containing keywords separately in biased and non-biased data according to Bert with their $F$ scores. It reveals that keywords are distributed in more concentrated paragraphs in biased data than that of in non-biased while sentence distribution of keywords may have little effect on the model performance.

5 Results and Analysis
Table 2 reveals a significantly different lexical choice in options between biased and unbiased data points for Bert. Right option tokens in biased dataset tend to be more prejudiced with higher $p$ scores and frequency variation, compared to non-biased data with more diverse vocabulary. Consequently, model may utilize such statistical cues for answering beyond understanding the passage. For example, “，” (a comma signal in Chinese characters usually used to express a parallel relationship) may be learned by model as a clue for right options. We infer that unbiased data should avoid repetitive and unvaried lexical choices in right option and reduce vocabulary differences with distracters. Table 3 illustrates that for Bert, sentence position distribution of annotated keywords has no obvious difference between two subsets ($P=0.441 > 0.05$), while keywords’ paragraph distribution differs in the performance gap ($P=0.002 < 0.005$). In other words, token-level supporting facts labeled by human are located in more concentrated paragraphs in biased samples with smaller average number of paragraphs containing keywords. This may due to the lack of considering about paragraph-level features in pre-train task designs. A more challenging MRC dataset can detect model reading comprehension level in terms of whole passage with complex text structure or more paragraphs.

6 Conclusion
In this paper, we construct a reading comprehension dataset HICMR with high-quality complex reasoning multi-choice questions and manually labelled supporting relevant facts in context, based on which we propose to identify biased samples with comprehensive consideration of human and model results. Our experiments reveal that baseline models behave differently from human when replacing full contexts with annotated keywords in the inputs, and Bert has an outstanding capability to capture the biases. We further explore the differences between biased and unbiased data in terms of lexical choice in options and evidence span distribution in passages. These results show that baseline models’ MRC capabilities may be overestimated due to biases or shortcuts in the datasets and there is still a long way to equip neural networks with higher quality and more challenging unbiased questions. One possible idea is to avoid high-frequency words or lexical choice preference in options, and employ consistent vocabulary among distracters and answer option. More complex paragraph structure would also be another suggestion to detect models’ reading comprehension abilities. We hope this work can inspire more researches in the future to adopt similar split method and evaluation scheme for MRC model evaluation.
References


Appendices

A Data Selection Criteria

- keep passages with longer length and multiple paragraphs.
- keep questions with four options and only one of them is right.
- remove options with apparent length bias, i.e. three short distracters and one longest answer option or vice versa.

B Annotation Procedure and Subjects Selection

B.1 Annotation Procedure

Step 1: Annotation preparation. Participants were trained on five exercise instances similar to experiment data, through which they become familiar with the task flow, annotation guideline and reading materials.

Step 2: Collaborative annotation. In view of plausibility, we split the task into two phases by one week. In phase one, each annotator is asked to finish 100 instances by reading a question and its corresponding passage (without options) and labeling up to 15 tokens that were relevant to answering the question. The number of labeled tokens is decided through pilot trial by authors considering average passage length. In phase two, annotators need to answer 200 questions, 100 of which were randomly mixed by the others’ annotation. They were presented with questions, options and masked passage where token not being marked was replaced with “_” and encouraged to select the right option by salary.

Step 3: Reliability monitoring. To ensure faithfulness, four unanswerable questions were mixed into experiment data to monitor cheating if participants acquired high accuracy including such data.

B.2 Subjects Selection

Participants should meet the following requirements:

- Chinese native speaker undergraduates.

- College Entrance Examination Chinese scores.

- No visual impairment.

- To avoid noise from age and gender, we set roughly equal number of male and female and the age from 18 to 30.

C Baseline Models

FastText It predicts probability of each option being right independently by encoding sentences as a bag of n-grams (Joulin et al., 2017). The option with the highest score is treated as the prediction for multiple-choice tasks. We employ the model in python library \(^1\) and keep the default hyperparameters settings.

Co-Matching It is a Bi-LSTM-based model and has reached promising results on RACE (Wang et al., 2018). It takes a question and its answer option as input sequences and learns to predict whether or not they match a given context. To keep it comparable, we use HanLP for Chinese word segmentation and the 300-dimensional Chinese word embeddings from (Li et al., 2018) as in C3.

Chinese Bert-Base We also apply the fine-tuning framework with a pre-trained language model Chinese Bert-Base from website, which has achieved impressive performance on MRC tasks (Devlin et al., 2018). For fine-tuning, we set batch size, learning rate, and maximal sequence length to 24, \(2 \times 10^{-5}\) and 512 respectively as they are in C3, and use default values for the other hyperparameters as in (Devlin et al., 2018).

D Biased Examples

---

1https://github.com/facebookresearch/fastText
There are so many flowers, so many flowers, but the crape myrtle is seldom mentioned, because it is so small, so it is spoken, insignificant. But she still has a forest grandness among feelings, flowers as beautiful. I like yang dynasty poet Yuan Mei's poem "eaves very much. The day is not everything, you just need to rely. Moss flowers as small as man, also like pony open. The poet described the moss grow quickly outside a way, but still grow green to show their youth. Where does youth come from? It comes from the strong vitality of moss, and its strong vitality, it breaks through difficulties and realizes the brilliance of youth. The crape does not bloom, but she does bloom, like the pony. And both are humble and proud. At this point, as long as you carefully observe, you will find that these insignificant moss, was so imposing. No matter on broken walls or cliffs, other plants could not settle down, only the moss from the cracks in the soil and stone, spreading green everywhere, exceeding in the rising spring breeze harder than any strange emblem. Father said, moss also has some poetic, its name is Qiong, also known as green insects, flowers with moss fell, the world will be full of spring. On the stage of moss, moss seems to have missed the Book of Songs, but caught up with the good times of Tang poetry and Song era, and also melted into the complicated affairs of Ming and Qing Dynasties. Small court spring all, green and hamamelis, moss always seems not to see the sun, only in the sad and miserable transgression growth. At this time, if you do not see moss, must be a pity. The world without moss is lonely.

Figure 1: Right answer:A, model predict using context:D, model predict using keywords:A

Figure 2: Right answer:D, model predict using context:C, model predict using keywords:D