Models can use keywords to answer questions that human cannot

Anonymous ACL submission

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.

	C3M test set	Our test set	Human Answerable set	Human Unanswerable set
Avg./Max. document length (in char)	180.2/1274	457/878	-	-
Avg./Max. question length (in char)	13.5/57	12.8/25	-	-
Avg./Max. option length (in char)	6.5/45	7.3/32	-	-
Single sent/Multiple sent/Independent	50.7/47.0/2.3	33.4/66.6/0	-	-
fastText	0.445	0.395 | 0.36	0.353	0.42
Co-matching	0.480	0.40 | 0.37	0.26	0.54
BERT	0.646	0.493 | 0.532	0.433	0.66
Human	0.933	0.78 | 0.72	0.88	0.445

Table 1: Statistics and reading comprehension accuracy of models and human on four datasets.

Agrawal et al. (2016) studied the behavior of models by variable length of the first question tokens in the field of visual QA. **?** stated that current task-oriented approaches in MRC typically develop a system and evaluate it on some specific datasets, resulting in lacks of generality but achieving extraordinary performance for that particular dataset. One of goals in this study motivated by these results was to identify biases exist in the current MRC datasets in more comprehensive manner.

Wiegreffe and Marasovic (2021) concluded three types of explanations including highlights, free-text and structured explanations. Inoue et al. (2020) 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 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).

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 HICRMC 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 HICRMC 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^1_{biased} =UH∩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^2_{biased} =UMC∩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^1_{biased}∪D^2_{biased}. 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∩UMK.

D_{biased}	$D_{non-biased}$
善| 0.54 | 24	和| 0.39 | 28
命| 0.5 | 24	这| 0.37 | 42
和| 0.49 | 33	而| 0.37 | 35
好| 0.45 | 31	，| 0.31 | 26
、| 0.44 | 107	理| 0.3 | 23
活| 0.44 | 64	类| 0.3 | 43
念| 0.44 | 23	学| 0.3 | 37
生| 0.43 | 141	事| 0.27 | 22
正| 0.44 | 52	文| 0.27 | 44
与| 0.44 | 59	者| 0.26 | 54

Table 2: Top 10 tokens that contribute to right options with more than 20 occurrences(token | p value | frequency).

	# of paragraphs containing keywords	# of sentences containing keywords
Biased	3.4 / 6	2.1 / 9
Non-biased	3.1 / 7	2.7 / 6
F	0.678	9.383
P-value	0.411	0.002
F crit	3.888	3.888

Table 3: Number (Avg./Max.) of sentences/paragraphs containing annotated keywords and significance test.

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 $l \in \{right, wrong\}$ given token t, where p(l/t)=count(t,l)/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

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Austin, Texas. Association for Computational Linguistics.

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

Lynette Hirschman, Marc Light, Eric Breck, and John D. Burger. 1999. Deep read: A reading comprehension system. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 325–332, College Park, Maryland, USA. Association for Computational Linguistics.

Xanh Ho, Anh-Khoa Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. pages 6609–6625.

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.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. pages 427–431.

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.

H. J. Levesque. 2014. On our best behaviour. *Artificial Intelligence*, 212:27–35.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on Chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143, Melbourne, Australia. Association for Computational Linguistics.

Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. 2018. Efficient and robust question answering from minimal context over documents. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1725–1735, Melbourne, Australia. Association for Computational Linguistics.

Praveen Paritosh. 2020. Achieving data excellence.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Durme. 2018. Hypothesis only baselines in natural language inference.

Roy Schwartz, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah Smith. 2017. Story cloze task: Uw nlp system. pages 52–55.

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.

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*.

Hongye Tan, Xiaoyue Wang, Yu Ji, Ru Li, Xiaoli Li, Zhiwei Hu, Yunxiao Zhao, and Xiaoqi Han. 2021. GCRC: A new challenging MRC dataset from Gaokao Chinese for explainable evaluation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1319–1330, Online. Association for Computational Linguistics.

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.

Fan Yang, Mengnan Du, and Xia Hu. 2019. Evaluating explanation without ground truth in interpretable machine learning. *CoRR*, abs/1907.06831.

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. *CoRR*, abs/2002.04326.

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, 2×10^{-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

[1] https://github.com/facebookresearch/fastText

	In Chinese	In English (by Google Translate)
context	百花繁，万花灿，唯有苔草很少被人提及，因为它实在微小，可以说是微不足道。但她依然有着茂林一般的风情、百花一样的美丽。我非常喜欢清代诗人袁枚那首《苔》诗：白日不到处，青春恰自来。苔花如米小，也学牡丹开。诗人笔下的青苔生长环境是很恶劣的，可它依然长出绿意来，展现出自己的青春。青春从何处来？它从苔草旺盛的生命力中来，它凭着坚强的活力，冲破困境，焕发青春的光彩。苔草是不会开花的，但她也学牡丹开，既是谦逊，也是骄傲。此时，只要你细心观察，就会发现这些微不足道的青苔，竟是如此有气势。无论是断墙残垣，还是悬崖绝壁之上，其它植物都无法落脚，唯有青苔从墙缝里、石缝隙中奋力拱出，四处蔓延着绿意，在荡漾的春风中记录着比石头还硬的倔强。父亲说，青苔也有一些诗意的名字，她叫绮线，也称呼为绿衣元宝，百花有青苔衬托，人世间才会春色满园。在岁月的戏台上，青苔似乎错过了《诗经》，却赶上了唐诗宋词的好时光，也融进了明清纷繁的花事。小庭春老，碧砌红萱草，青苔似乎总是见不到阳光，只在凄凄惨惨中顽强地生长着。此时，如果你没有见青苔，一定是遗憾的；没有青苔的世界，也是寂寞的。	There are so many flowers, so many flowers, but the carex plant is seldom mentioned, because it is so small, so to speak, insignificant. But she still has a forest general amorous feelings, flowers as beautiful. I like qing dynasty poet Yuan Mei's poem "Moss" very much: The day is not everywhere, youth just come. Moss flowers as small as rice, also like peony open. The poet described the moss growth environment is very bad, but it still grow green to show their youth. Where does youth come from? It comes from the strong vitality of carex, and with its strong vitality, it breaks through difficulties and radiates the brilliance of youth. The carex does not blossom, but she does blossom like the peony, and is both 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 wall and stone, spreading green everywhere, recording in the rippling spring breeze harder than stone stubborn. Father said, moss also has some poetic names, her name is Qixian, also known as green ingot, flowers with moss foil, the world will be full of spring. On the stage of time, moss seems to have missed the Book of Songs, but caught up with the good times of Tang poetry and Song ci, and also melted into the complicated affairs of Ming and Qing Dynasties. Small court spring old, green red hemerocallis, moss always seems not to see the sun, only in the sad and miserable tenacious growth. At this time, if you do not see moss, must be a pity; The world without moss is lonely.
question	以下选项中对本文的主旨思想理解最为准确的一项是?	Which one is the most suitable main idea of the passage?
options	A. 赞美青苔的顽强和倔强 B. 陈述青苔对春天的点缀 C. 青苔跟其他植物一样美丽 D. 青苔没有被写进《诗经》是值得遗憾的	A. Praise the moss tenacious and stubborn B. Stating the ornament of moss to spring C. Moss is as beautiful as any other plant D. It is a pity that moss was not written into the Book of Songs
Annotated keywords	_花______苔草__________微小_________但_________风情___一样_美丽____________苔 ______________________ ____________青春_青春____________生命力______坚强___ 冲破困境______________________谦逊___骄傲______细 ____________青苔______气势 ______________________________缝 ________________________________倔强 ______________________ _____岁月 ______________________ ________________________顽强_生长________没有_青苔 ___遗憾__没有青苔_____寂寞__	--

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

	In Chinese	In English (by Google Translate)
context	1月11日7时35分，河南鹤壁市贾家村小学一男一女两名小学生上学时，为走近道，在通过鹤壁北站编车车辆时，从车底下钻过，却没注意对面有火车驶来的信号。这时，旁边轨道上的列车正向两名小学生疾驶而来，男学生由于跑得快，跌倒在铁轨外，而女学生早已吓得晕在轨道上。就在这万分危急的时刻，正在旁边执行任务的该站检车员陈宝昌，奋不顾身冲上前去，猛地扑在女学生身上。此时，陈宝昌把小女孩儿抢出铁轨已不可能了，使用自己的身体将女学生紧紧压在轨道中间。从陈宝昌及女学生身上飞驶而过的列车剐破了陈宝昌身上的棉衣，幸运的是两个人都未受伤。但那个男生由于跌倒在铁轨外，左脚脚趾被列车车轮轧掉。陈宝昌把早已吓得面色苍白的女学生背出轨道，又和同事一起，把受伤的男生送往医院抢救。今年27岁的陈宝昌，平时工作努力，乐于助人，曾连续3年获得郑州铁路分局新乡车辆段“优秀团支部书记”称号。	At 7:35 am on January 11, a boy and a girl from Jiajia Village Primary School in Hebi city, Henan province, went under a train to get near the north Hebi Railway Station, but did not pay attention to the signal of an approaching train. At this time, the train on the side of the track is driving two pupils, male students as a result of running fast, fall outside the track, and female students have been scared dizzy on the track. In this extremely critical moment, is next to the task of the station inspection car member Chen Baochang, regardless of personal danger rushed forward, suddenly on the female students. At this point, It was impossible for Chen baochang to snatch the girl off the track, so he used his body to press the girl in the middle of the track. From Chen Baochang and female students who flew by the train cut Chen Baochang's body cotton-padded clothes, fortunately, both people were not injured. But the boy fell off the track and the toe of his left foot was crushed by the wheel of the train. Chen Baochang had been scared pale female students back out of the track, and colleagues together, the injured boy to the hospital for rescue. Chen Baochang, 27 years old this year, usually works hard and is ready to help others. He has won the title of "Excellent League Branch Secretary" of Xinxiang Rolling Stock Section of Zhengzhou Railway Sub-bureau for three consecutive years.
question	两名小学生上学时，为什么要从火车车厢底下钻过去?	Why did two schoolchildren go under a train carriage on their way to school?
options	A. 觉得好玩儿 B. 看见了陈宝昌 C. 上课时间快到了 D. 为了方便、省时间	Find it amusing B. They saw Chen Baochang C. It's almost time for class D. For convenience and time saving
Annotated keywords	______________________上学__为走近道 ______________钻过__没注意___火车驶来_信号_续 ______________________	--

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