DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications

Wei He, Kai Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, Haifeng Wang
Baidu Inc., Beijing, China
{hewei06, liukai20, lvyajuan, zhaoshiqi, xiaoxinyan, liuyuan04, wangyizhong01, wu_hua, sheqiaoqiao, liuxuan, wutian, wanghaifeng}@baidu.com

Abstract

In this paper, we introduce DuReader, a new large-scale, open-domain Chinese machine reading comprehension (MRC) dataset, aiming to tackle real-world MRC problems. In comparison to prior datasets, DuReader has the following characteristics: (a) the questions and the documents are all extracted from real application data, and the answers are human generated; (b) it provides rich annotations for question types, especially yes-no and opinion questions, which take a large proportion in real users' questions but have not been well studied before; (c) it provides multiple answers for each question. The first release of DuReader contains 200k questions, 1,000k documents, and 420k answers, which, to the best of our knowledge, is the largest Chinese MRC dataset so far. Experimental results show there exists big gap between the state-of-the-art baseline systems and human performance, which indicates DuReader is a challenging dataset that deserves future study. The dataset and the code of the baseline systems are publicly available now[1].

1 Introduction

For human beings, reading comprehension is a basic ability to acquire knowledge. We believe it is one of the crucial abilities machine has to have to acquire knowledge through reading the whole web and answer open domain questions. Such an ability is considered to be of great value for next-generation search engines and intelligent agent products. However, Machine Reading Comprehension (MRC) is an extremely challenging work since it involves several difficult tasks such as comprehension, inference and summarization.

Recently, several MRC datasets have been released, greatly inspiring the research in this field. A series of neural network models, such as Match-LSTM (Wang and Jiang, 2017), BiDAF (Seo et al., 2016), R-net (Wang et al., 2017), have been proposed, achieving promising results on a variety of MRC evaluation tasks.

However, most existing MRC datasets have some limitations due to their synthetic data, simplified tasks or constrained domains. Therefore, studies on these datasets are different from real-world comprehension tasks. In detail, cloze-style MRC (Hermann et al., 2015; Hill et al., 2015; Cui et al., 2016) simplifies the task into word prediction on hole-digging synthesis data. Multiple-choice MRC (Lai et al., 2017) tests comprehension ability via option selection on examination data. Question answering based MRC (Trischler et al., 2017; Rajpurkar et al., 2016; Joshi et al., 2017) usually casts reading comprehension as the prediction of span in a news article, a Wikipedia entry or other documents for a given question. Although such kinds of simplifications and constraints facilitate the data construction and the model design, they bring some undesired problems. By analyzing questions real users submitted to Baidu search engine, we found that current datasets cover only some types of questions, leaving other types, such as opinion questions and complex description questions, not well studied. Furthermore, recent studies (Chen et al., 2016; Jia and Liang, 2017) have shown that current MRC models could achieve high performance on many of these datasets with limited comprehending or

[1] The DuReader dataset is available at `http://ai.baidu.com/broad/subordinate?dataset=dureader`. The code of the baseline systems is open sourced at `https://github.com/baidu/DuReader`

Dataset	lang.	#queries	#doc.	query source	doc. source	answer
CNN/Daily Mail	ENG	1.4M	300k	Cloze	News	Fill in entity
HLF-RC	CHN	100k	28k	Cloze	Fairy/News	Fill in word
RACE	ENG	870k	50k	English exam	English exam	Multi. choices
NewsQA	ENG	100k	10k	Crowdsourced	CNN	Span of words
SQuAD	ENG	100k	536	Crowdsourced	Wiki.	Span of words
TrivaQA	ENG	40k	660k	Trivia websites	Wiki./Web doc.	Span/substring of words
MS-MARCO	ENG	100k	200k[2]	User logs	Web doc.	Summary by human
DuReader	CHN	200k	1000k	User logs	Web doc./CQA	Summary by human

Table 1: Comparison of some properties of existing datasets[3] vs. DuReader.

	Fact	**Opinion**
Entity	iphone哪天发布 On which day will iphone be released	2017最好看的十部电影 Top 10 movies of 2017
Description	消防车为什么是红的 Why are firetrucks red	丰田卡罗拉怎么样 How is Toyota Carola
YesNo	39.5度算高烧吗 Is 39.5 degree a high fever	学围棋能开发智力吗 Does learning to play go improve intelligence

Table 2: Examples for question types from two views.

inferring.

Therefore, it is necessary to build real-world reading comprehension datasets in open domain. An English dataset, MS-MARCO (Nguyen et al., 2016) was released under this consideration, in which the questions and documents were collected from search engine, and answers were generated by human annotators. In this paper, we propose DuReader, a new large-scale and human annotated MRC dataset in Chinese language, aiming to tackle real-world MRC problems. Besides its merits that the questions are open-domain and extracted from real application data, DuReader has the following characteristics compared to previous datasets.

1. DuReader provides rich annotations for question types. In particular, DuReader annotates yes-no and opinion questions that take a large proportion in real user's questions but have not been well studied before. Answering opinion questions usually requires inference and summarization of multiple evidences, which are challenging even for human.

2. DuReader collects documents from the search results of Baidu search engine and from a question answering community site Baidu Zhidao. All the contents in Zhidao are contributed by its users, making its documents more colloquial and different from common web pages.

3. DuReader provides multiple answers for each question. Most previous datasets have only one answer for each question while in real world, one question may have one or several answers depending on what the question is. DuReader can reflect real-world applications more than other datasets.

The first release of DuReader contains 200k questions, 1M documents and more than 420k human-summarized answers. To the best of our knowledge, DuReader is the largest Chinese MRC dataset so far. The comparison of some key properties of DuReader and the existing datasets is shown in Table 1.

We implemented two state-of-the-art MRC models, i.e., Match-LSTM (Wang and Jiang, 2017) and BiDAF (Seo et al., 2016) on DuReader. We find that the performances of these models are far inferior to human, which suggests that there is a large room for researchers to improve the MRC models on DuReader dataset.

2 Analysis of Questions in Search Engine

In this section, we analyze the distribution of questions in Baidu Search data. We randomly sample 1,000 questions from one day's search log,

[2]The number is calculated on unique documents.

[3]CNN/Daily Mail (Hermann et al., 2015), HLF-RC (Cui et al., 2016), RACE (Lai et al., 2017), NewsQA (Trischler et al., 2017), SQuAD (Rajpurkar et al., 2016), TrivaQA (Joshi et al., 2017), MS-MARCO (Nguyen et al., 2016)

	Fact	Opinion	Total
Entity	23.4%	8.5%	31.9%
Description	34.6%	17.8%	52.5%
YesNo	8.2%	7.5%	15.6%
Total	66.2%	33.8%	100.0%

Table 3: Distribution of question types in Baidu search data.

and then manually annotate the questions from two different views. From the view of the answer type that a question belongs to, we classify the questions into three kinds: *Entity*, *Description* and *YesNo*. For *Entity* questions, the answers are expected to be a single entity or a list of entities. For *Description* questions, the answers are usually multi-sentence summaries. This kind of questions contain how/why questions, questions of comparing the functions of two or more objects, questions about inquiring the merits/demerits of a goods, etc. As for *YesNo* questions, the answers are expected to be an affirmative or negative answers with supportive evidences.

After a deep investigation into the questions, we found that whichever answer type a question belongs to, it can be further classified into *Fact* or *Opinion*, depending on whether it is about a fact or an opinion[4]. Table 2 shows some examples.

For each question in the sampled data, we label it from two views: one is the answer type it belongs to, the other is whether it is about fact or opinion. In this way, the questions can be classified into six classes. The distribution of the questions in the sample data is shown in Table 3.

From the distribution, some interesting phenomena are observed:

1. The *Entity-Fact* questions, also known as factoid questions that have been widely studied in previous work, account only for 23.4%.

2. Over half of the questions (52.5%) are Description questions. Previous studies mostly focus on Description-Fact questions.

3. *YesNo* questions accounts for 15.6%, with one half about fact, another half about opinion.

4. More than one-third of questions are *Opinion* questions, seldom addressed in the previous research.

To the best of our knowledge, it is the first time to analyze the MRC dataset from two different views. Some of the question types have been widely studied in previous work while others, especially *YesNo* questions and *Opinion* questions, are expecting more attention from researchers. Hopefully, our datasets can promote further researches on them.

3 DuReader Dataset

In this section, we will introduce the data collection and annotation process of DuReader. The DuReader dataset can be considered as a set of quadruples of $\{q, t, D, A\}$, which are defined as: (a) the question q; (b) the question type t; (c) the relevant document set D=d_1,d_2,...,$d_{|D|}$, which contains $|D|$ documents; (d) reference answers set A which is generated by human annotators.

3.1 Data Collection and Annotation

3.1.1 Data Collection

To collect questions for DuReader, we first randomly sampled frequently occurring queries from Baidu search engine query logs. Questions were filtered from the queries using a binary classifier, which were then double-checked by human annotators. 200K questions are reserved in this release.

We then collected relevant documents for the questions. Two sources were explored, i.e., search results of Baidu search engine, and Baidu Zhidao[5]. In detail, the 200K questions were divided into two sets. For the first half, we searched each question in Baidu search engine and kept top-5 search results, while for the second half, we searched each question in Zhidao's site search and also kept top-5 results. The reason why we use Zhidao as a source of relevant documents, is that the User Generated Content (UGC) nature of Zhidao makes its documents different from random web pages on the Internet. Especially, for the opinion questions, there are more answers in Zhidao.

[4]According to the definition of opinion in wikipedia (`https://en.wikipedia.org/wiki/Opinion`), an opinion is a judgment, viewpoint, or statement that is not conclusive. It may deal with subjective matters in which there is no conclusive finding. What distinguishes fact from opinion is that facts are more likely to be verifiable, i.e. can be agreed to by the consensus of experts.

[5]Zhidao (`https://zhidao.baidu.com`) is the largest Chinese community-based question answering (CQA) site in the world.

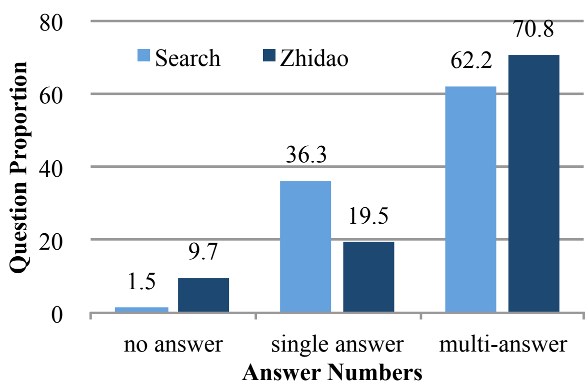

Figure 1: Distribution of answer numbers per question in DuReader.

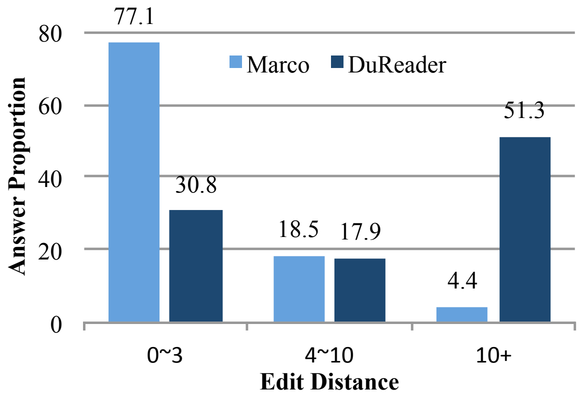

Figure 2: Distribution of edit distance between answers and original documents.

For each document, we extracted the title and main contents, which were then word-segmented using the open API of the Baidu AI platform[6].

3.1.2 Question Type Annotation

According to the two-dimension question types introduced in Section 2, the annotators were asked to label each question in a two-pass manner. In the first pass, the annotators classified all the questions into three types: *Entity*, *Description* and *YesNo* questions. And in the second pass, the annotators labeled each question as either *Fact* or *Opinion*. The distribution of questions of different types in DuReader is shown in Table 4. Note that the distribution of question types in DuReader (Table 4) is different from that in Baidu Search (Table 3). This is mainly because Table 4 is type-based statistics, since we keep only one instance in the dataset for same questions, while the statistics for Table 3 is frequency-based.

3.1.3 Answer Annotation

For the answer annotation, we employed crowdsourcing workers to generate answers for each question based on the relevant documents. Specifically, each question and its relevant documents were shown to an annotator. He/she was asked to generate answers in his/her own words by reading and summarizing the documents. If more than one answer can be found in the relevant documents, the annotator was required to write down all the answers. Those answers that are very similar to each other were merged into only one. The answers were pointed-checked to guarantee that the accuracy is high enough.

[6]http://ai.baidu.com/tech/nlp/lexical

	Fact	**Opinion**	**Total**
Entity	14.4%	13.8%	28.2%
Description	42.8%	21.0%	63.8%
YesNo	2.9%	5.1%	8.0%
Total	60.1%	39.9%	100.0%

Table 4: Distribution of question types in DuReader.

Specifically, for the *Entity* questions, the answers include both the entities and the sentences containing them. For the *YesNo* questions, the answers include the opinion types (*Yes*, *No* or *Depend*) as well as the supporting sentences. (See the last example in Table 9)

3.2 Data Analyzing

Statistics on length. On average, each question and answer has 4.8 and 69.6 words respectively. The average length of the documents is 396.0 words, which is about 5 times longer than those in MS-MARCO. The reason is that we kept the full body of each relevant document whereas MS-MARCO only use a certain paragraph.

Statistics on answer numbers. Figure 1 shows the statistics over number of answers. We can see that 1.5% of Baidu Search questions have zero answers, but this number increases to 9.7% for Baidu Zhidao. Meanwhile, Baidu Zhidao has a larger proportion of multi-answer questions than Baidu Search (70.8% vs. 62.2%), which may be explained as the UGC answers are more subjective and cause more diversity.

Difficulty analysis of DuReader. In order to understand the difficulty to answer the problem in DuReader, we calculate the distribution of minimum edit distance (ED) between the answers generated by human and the original documents[7]. The larger ED is, the more summarization and modification has been operated by the annotators, which requires more complex methods for modeling the problem. For the span-answer datasets, such as SQuAD (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2017) and TriviaQA (Joshi et al., 2017) the ED score should be zero, since the answers are all directly extracted from the document. Figure 2 shows the distribution of ED scores between the answers and documents in DuReader, which is compared with those in MS-MARCO. We can see that for 77.1% of answers in MS-MARCO the ED is below 3. In contrast, 51.3% of DuReader answers has a ED score over 10, which can be inferred that DuReader requires more complex techniques such as text summarization and generation.

4 Experiments

In this section, we implement MRC systems with two state-of-the-art models. BLEU (Papineni et al., 2002) and Rouge(Lin, 2004) are used as the basic evaluation metrics. Furthermore, with the rich annotations in our dataset, including the queries and answers of various types, we conduct comprehensive evaluations from different aspects.

4.1 Baseline Systems

We implement two typical state-of-the-art models as baseline systems.

Match-LSTM Match-LSTM is a widely used MRC model and has been well explored in recent studies (Wang and Jiang, 2017). To find an answer in the passage, it goes through the passage sequentially and dynamically aggregates the matching of an attention-weighted question representation to each token of the passage. Finally, an answer pointer layer is used to find an answer span in the passage.

BiDAF BiDAF is a promising MRC model, and its improved version has achieved the best single model performance on SQuAD dataset (Seo et al., 2016). It uses both context-to-question attention and question-to-context attention in order to highlight the important parts in both question and context. After that, the so-called attention flow layer is used to fuse all useful information in order to get a vector representation for each position.

To set up, we randomly initialize the word embeddings with a dimension of 300 and set the hidden vector size as 150 for all layers. We use the Adam algorithm (Kingma and Ba, 2014) to train both models with an initial learning rate of 0.001 and a batch size of 32. Since for every question there may be multiple corresponding passages. To improve training and testing efficiency, a simple heuristic strategy is employed to select a representative paragraph from each passage. This paragraph is supposed to be the one that achieves the highest recall score when compared against the annotated answers during training. While for testing, since the answers are not available, we compute the recall score against the question instead.

4.2 Results and Analysis

We evaluate the reading comprehension task via character-level BLEU-4 (Papineni et al., 2002) and Rouge-L (Lin, 2004), which are widely used for evaluating the quality of language generation. The experimental results on test set are shown in Table 5. For comparison, we also evaluate the Selected Paragraph system, which directly selects the paragraph that achieving the highest recall score as answer. And we assess human performance by involving a new annotator to annotate on the test data and treat his first answer as the prediction.

The results demonstrate that current reading comprehension models can achieve an impressive improvement compared with the selected paragraph baseline, which approves the effectiveness of these models. However, there is still a large performance gap between these models and human. An interesting discovery comes from the comparison between results on Search and Zhidao data. We find that the reading comprehension models get much higher score on Zhidao data. This shows that it is much harder for the models to comprehend open-domain web articles than to find answers in passages from a question answering community. In contrast, the performance of human beings on these two datasets shows little difference, which suggests that human's reading skill is more stable on different types of documents.

As described in Section 4.1, the representative paragraph of each passage is selected based on the query during testing. To analyze the effect of the strategy and obtain the upper bound of the base-

[7] Here ED is the minimum edit distance between the answer and any consecutive span in the document.

Systems	Search		Zhidao		All	
	BLEU-4%	Rouge-L%	BLEU-4%	Rouge-L%	BLEU-4%	Rouge-L%
Selected Paragraph	15.8	22.6	16.5	38.3	16.4	30.2
Match-LSTM	23.1	31.2	42.5	48.0	31.9	39.2
BiDAF	23.1	31.1	42.2	47.5	31.8	39.0
Human	55.1	54.4	57.1	60.7	56.1	57.4

Table 5: Performance of typical MRC systems on the DuReader dataset.

	BLEU-4%	Rouge-L%
Gold Paragraph	31.7	61.3
Match-LSTM	46.3	52.4
BiDAF	46.3	51.8

Table 6: Model performance with gold paragraph.

line models, we re-evaluate our systems on the gold paragraphs, each of which is selected if its recall score against the annotated answers is the highest. Comparing Table 6 with Table 5, we can see that the use of gold paragraphs could significantly boosts the overall performance. Moreover, directly using the gold paragraph can obtain a very high Rouge-L score, which is as expected because each gold paragraph is selected based on recall, which is relevant to Rouge-L. Though, we still find that the baseline models can get much better performance with respect to BLEU, which means the models have learned to refine the answers. The experiment shows that paragraph selection is a crucial problem to solve in real applications, while most current MRC datasets suppose to find the answer in a given small passage. Thus, DuReader provides full body text of evidence document to stimulate research in real-world setting.

To gain more insight into the characteristics of our dataset, we report the performance across different question types in Table 7. We can see that both the models and human achieve relatively good performance on description questions, while *YesNo* questions seem to be the hardest to model. We consider that description questions are usually answered with long text on the same topic. This is preferred by BLEU or Rouge. However, the answers of *YesNo* questions are relatively short, which could be a simple *Yes* or *No* in some cases. Even more interesting is, the answers of some *YesNo* questions are quite subjective and some may even be contradictory based on the evidence collected from different passages. Therefore, even the human annotators cannot reach a high level of agreement for these questions.

4.3 Opinion-aware Evaluation

Considering the characteristics of *YesNo* questions, we found that it's not suitable to directly use BLEU or Rouge to evaluate the performance on these questions, because these metrics could not reflect the agreement between answers. For example, two contradictory answers like "You can do it" and "You can't do it" get high agreement scores with these metrics. A natural idea is to formulate this subtask as a classification problem. However, as described in Section 3, multiple different judgments could be made based on the evidence collected from different passages, especially when the question is of opinion type. In real-world settings, we definitely don't want a smart model to give an arbitrary answer for such questions as *Yes* or *No*.

To tackle this, we propose a novel opinion-aware evaluation method that requires the evaluated system to not only output an answer in natural language, but also give it an opinion label. We also have the annotators provide the opinion label for each answer they generated. In such cases, every answer is paired with an opinion label (*Yes*, *No* or *Depend*) so that we can categorize the answers by their labels. Finally, the predicted answers are evaluated via Blue or Rouge against only the reference answers with the same opinion label. By using this opinion-aware evaluation method, a model that can predict a good answer in natural language and give it an opinion label correctly will get a higher score.

In order to classify the answers into different opinion polarities, we add a classifier. We slightly change the Match-LSTM model, in which the final pointer network layer is replaced with a fully connected layer. This classifier is trained with the gold answers and their corresponding opinion labels. We compare a reading comprehension system equipped with such an opinion classifier with

Question type	Description		Entity		YesNo	
	BLEU-4%	Rouge-L%	BLEU-4%	Rouge-L%	BLEU-4%	Rouge-L%
Match-LSTM	32.8	40.0	29.5	38.5	5.9	7.2
BiDAF	32.6	39.7	29.8	38.4	5.5	7.5
Human	58.1	58.0	44.6	52.0	56.2	57.4

Table 7: Performance on various question types.

	Fact		Opinion	
	BLEU-4%	Rouge-L%	BLEU-4%	Rouge-L%
Opinion-unaware	6.3	8.3	5.0	7.1
Opinion-aware	12.0	13.9	8.0	8.9

Table 8: Performance of opinion-aware model on *YesNo* questions.

a pure reading comprehension system without it, and the results are demonstrated in Table 8. We can see that doing opinion classification does help under our evaluation method. Also, classifying the answers correctly is much harder for the questions of opinion type than for those of fact type.

4.4 Discussion

As shown in the experiments, the current state-of-the-art models still underperform human beings by a large margin on our dataset. There is considerable room for improvement on several directions.

First, the state-of-the-art models formulate reading comprehension as a span selection task. However, as shown in DuReader dataset, human beings actually summarize answers with their own comprehension. How to summarize or generate the answers deserves more research. Current methods employ a simple paragraph selection strategy, which results in great degradation of comprehension accuracy as compared to gold paragraph's performance. It is necessary to design novel and efficient whole-document representation models for the real-world MRC problem.

Second, there are some new features in our dataset that have not been extensively studied before, such as yes-no questions and opinion questions requiring multi-document MRC. New methods are needed for opinion recognition, cross-sentence reasoning, and multi-document summarization. Hopefully, DuReader's rich annotations would be useful for study of these potential directions.

Third, as the first release of the dataset, it is far from perfection and it leaves much room for improvement. For example, we annotate only opinion tags for yes-no questions, we will also annotate opinion tags for description and entity questions. We would like to gather feedback from the research community to improve DuReader continually.

Overall it is necessary to propose new algorithms and models to tackle with real-world reading comprehension problems. We hope that the DuReader dataset would be a good start for facilitating the MRC research.

5 Conclusion and Future Work

We introduce DuReader, a new Chinese large-scale open domain dataset for machine reading comprehension. Different from exiting Chinese MRC datasets, DuReader contains questions and possible answers from real-world applications, with the aim to promote MRC research in real-world setting. In particular, DuReader contains rich annotations of questions, documents and answers. It is the first time to annotate the questions from two different views, among which yes-no and opinion questions account for a large proportion but have not been well studied yet. For each question, we provide documents coming from both Baidu Search and Baidu Zhidao, and multi-answers with supporting evidence, possible entities and opinions labelled. Hopefully, these annotations could help in facilitating MRC research. Preliminary experimental results show that there exists a significant gap between the performances of state-of-the-art models and that of humans on this dataset.

In future work, we will steadily update our dataset by enlarging the size and enriching the annotations based on feedbacks from the community. We expect DuReader will be a valuable resource to

Question	学士服颜色/ What are the colors of academic dresses?
Question Type	*Entity-Fact*
Answer 1	**[绿色, 灰色, 黄色, 粉色]**：农学学士服绿色，理学学士服灰色，工学学士服黄色，管理学学士服灰色，法学学士服粉色，文学学士服粉色，经济学学士服灰色。/ **[green, gray, yellow, pink]** Green for Bachelor of Agriculture, gray for Bachelor of Science, yellow for Bachelor of Engineering, gray for Bachelor of Management, pink for Bachelor of Law, pink for Bachelor of Art, gray for Bachelor of Economics
Document 1	农学学士服绿色，理学学士服灰色，...，确定为文、理、工、农、医、军事六大类，与此相应的饰边颜色为粉、灰、黄、绿、白、红六种颜色。
...	
Document 5	学士服是学士学位获得者在学位授予仪式上穿戴的表示学位的正式礼服，...，男女生都应着深色皮鞋。
Question	迈腾和帕萨特哪个好? / Which car is better, MAGOTAN or PASSAT?
Question Type	*Description-Opinion*
Answer 1	迈腾稍微好点，具体看配置吧/ PASSAT may be better, but it actually depends on your custom options
Answer 2	两车性能基本一致，只是外形不同/ almost the same, just different in appearance
Document 1	虽然从审美的角度来看,帕萨特不见得有多落伍,但和采用了全新设计的迈腾相比...
...	
Document 5	迈腾稍微好点,但是迈腾那个变速箱真不如帕萨特。具体看配置吧,配置高的帕萨特感觉还是...
Question	智慧牙一定要拔吗/ Do I have to have my wisdom teeth removed
Question Type	*YesNo-Opinion*
Answer 1	**[Yes]**因为智齿很难清洁的原因，比一般的牙齿容易出现口腔问题，所以医生会建议拔掉/ **[Yes]** The wisdom teeth are difficult to clean, and cause more dental problems than normal teeth do, so doctors usually suggest to remove them
Answer 2	**[Depend]**智齿不一定非得拔掉，一般只拔出有症状表现的智齿，比如说经常引起发炎... / **[Depend]** Not always, only the bad wisdom teeth need to be removed, for example, the one often causes inflammation ...
Document 1	为什么要拔智齿？智齿好好的医生为什么要建议我拔掉?主要还是因为智齿很难清洁...
...	
Document 5	根据我多年的临床经验来说,智齿不一定非得拔掉.智齿阻生分好多种...

Table 9: Examples from DuReader dataset

the development of MRC technologies and applications.

References

[Chen et al.2016] Danqi Chen, Jason Bolton, and Christopher D Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2367.

[Cui et al.2016] Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. 2016. Consensus attention-based neural networks for chinese reading comprehension.

[Hermann et al.2015] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

[Hill et al.2015] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

[Jia and Liang2017] Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2011–2021.

[Joshi et al.2017] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *CoRR*.

[Kingma and Ba2014] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*.

[Lai et al.2017] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

[Lin2004] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

[Nguyen et al.2016] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

[Papineni et al.2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

[Rajpurkar et al.2016] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

[Seo et al.2016] Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *CoRR*.

[Trischler et al.2017] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200.

[Wang and Jiang2017] Shuohang Wang and Jing Jiang. 2017. Machine comprehension using match-lstm and answer pointer. In *ICLR*, pages 1–15.

[Wang et al.2017] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated Self-Matching Networks for Reading Comprehension and Question Answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198.