DuReader$_{\text{robust}}$: A Chinese Dataset Towards Evaluating the Robustness of Machine Reading Comprehension Models

Hongxuan Tang[1]* **Jing Liu**[2] **Hongyu Li**[2] **Yu Hong**[1] **Hua Wu**[2] **Haifeng Wang**[2]
[1]School of Computer Science and Technology, Soochow University, China
[2]Baidu Inc., Beijing, China
{hxtang01, tianxianer}@gmail.com
{liujing46, lihongyu04, wu_hua, wanghaifeng}@Baidu.com

Abstract

Machine Reading Comprehension (MRC) is a crucial and challenging task in natural language processing. Although several MRC models obtains human parity performance on several datasets, we find that these models are still far from robust. To comprehensively evaluate the robustness of MRC models, we create a Chinese dataset, namely DuReader$_{\text{robust}}$. It is designed to challenge MRC models from the following aspects: (1) over-sensitivity, (2) over-stability and (3) generalization. Most of previous work studies these problems by altering the inputs to unnatural texts. By contrast, the advantage of DuReader$_{\text{robust}}$ is that its questions and documents are natural texts. It presents the robustness challenges when applying MRC models to real-world applications. The experimental results show that MRC models based on the pre-trained language models perform much worse than human does on the robustness test set, although they perform as well as human on in-domain test set. Additionally, we analyze the behavior of existing models on the robustness test set, which might give suggestions for future model development. The dataset and codes are available at https://github.com/PaddlePaddle/Research/tree/master/NLP/DuReader-Robust-BASELINE

1 Introduction

Machine reading comprehension (MRC) requires machines to answer questions conditioned on understanding the given text, and it is an important and challenging task in natural language processing. In recent years, with the increasing availability of large-scale datasets (Rajpurkar et al., 2016; Nguyen et al., 2016) and the development of deep learning, MRC has achieved remarkable advancements (Seo et al., 2016; Wang and Jiang, 2016). In particularly, pre-trained language models (LM) (Devlin et al., 2018; Liu et al., 2019; Lan et al., 2020), have caused a stir in the MRC community and have presented new state-of-the-art results.

Although several neural models obtain high accuracy on several datasets, previous studies show that most of the neural models are not robust enough (Jia and Liang, 2017; Ribeiro et al., 2018b; Welbl et al., 2019). The robustness issues of MRC models may limit their usage in the real-world applications, e.g. search engines and dialogue systems, since such issues may lead to unexpected predictions on various inputs and hurt user experience. In this paper, we focus on the following three robustness issues.

(1) **Over-sensitivity issue**. By over-sensitivity, we mean that an MRC model provides different answers to paraphrase questions, when it should not. This suggests that the model is over-sensitive to the minor difference between the paraphrase questions. As shown in Table 1a, the model gives a correct answer to the original question, while it gives a different answer to the paraphrase question.

(2) **Over-stability issue**. By over-stability, we mean that an MRC model overly relies on spurious lexical patterns without language understanding. Hence, the model is bad at distinguishing a sentence containing ground-truth answers from a distracting sentence, that just has many words in common with the question. As shown in Table 1b, the model is confused by a distracting sentence, and gives a wrong prediction that locates in the distracting sentence.

(3) **Generalization issue**. Current models usually generalize well to in-domain test sets, yet perform poorly on out-of-domain test sets that have the distributions different from the training sets. As shown in Table 1c, the model that is trained on a training set with general domain, gives a wrong pre-

*This work was done while the first author was an intern at Baidu Inc.

Passage 史蒂夫·乔布斯，1955年2月24日生于美国加利福尼亚州旧金山，美国发明家、企业家、美国苹果公司联合创办人。2011年10月5日，因胰腺癌病逝，享年56岁…….	**Passage** Steve Jobs was born in San Francisco, California, USA on February 24, 1955. He is an American inventor, entrepreneur and co-founder of Apple Inc. On October 5, 2011, he died of pancreatic cancer at the age of 56 ...
Original Question 乔布斯几岁死 **Golden Answer**： 56岁 **Predicted Answer**： 56岁 (BERT-Base)	**Original Question** What was the age of Steve Jobs when he died? **Golden Answer** : 56 **Predicted Answer** : 56 (BERT-Base)
Paraphrase Question 乔布斯多大死的 **Golden Answer**： 56岁 **Predicted Answer**： 胰腺癌 (BERT-Base)	**Paraphrase Question** How old was Steve Jobs when he died? **Golden Answer** : 56 **Predicted Answer** : Pancreatic cancer (BERT-Base)

(a) An example to demonstrate the over-sensitivity issue of an MRC model. The model gives two different predictions to two paraphrased questions. Hence, we consider this model is over-sensitive.

Passage 包粽子的线以前人们认为是来自麻叶子,其实是棕榈树,粽子的音就来自棕叶子。	**Passage** The ropes of rice dumpling was previously thought to made with Folium Cannabis, but it was actually made with palm, and the pronunciation of rice dumpling came from palm as well.
Question 包粽子的线来自什么 **Golden Answer**： 棕榈树 **Predicted Answer**： 麻叶子 (BERT-Base)	**Question** What are the ropes of rice dumpling made with? **Golden Answer** : palm **predicted Answer** : Folium Cannabis (BERT-Base)

(b) An example to demonstrate the over-stability issue of an MRC model. The distracting sentence "The ropes of rice dumpling was previously thought to made with Folium Cannabis" has more words in common with the question "What are the ropes of rice dumpling made with?". The model relies on lexical patterns too much without language understanding, and we consider this model is over-stable.

Passage cos(2x)'=-sin(2x)*(2x)'=-2sin(2x) 属于复合函数的求导	**Passage** cos(2x)'=-sin(2x)*(2x)'=-2sin(2x) it belongs to derivation of compound functions
Question cos2x的导数是多少? **Golden Answer** : -2sin(2x) **Predicted Answer** : -sin(2x) (BERT-Base)	**Question** What is the derivative of cos2x? **Golden Answer** : -2sin(2x) **Predicted Answer** : -sin(2x) (BERT-Base)

(c) An example to demonstrate the generalization issue of an MRC model. The model that is trained on a training set with general domain, gives a wrong prediction to a math question, since the model lacks of the knowledge in math.

Table 1: The examples of over-sensitivity, over-stability and generalization issues.

diction to a math question, since the model lacks of the knowledge in math.

In previous work, the above issues have been studied separately. In this paper, we aim to create a dataset namely DuReader$_{\text{robust}}$ to comprehensively evaluate the three robustness issues of neural MRC models. Previous work studies these issues by altering the questions or the documents. Ribeiro et al. (2018b); Iyyer et al. (2018); Gan and Ng (2019) try to evaluate the over-sensitivity issue via paraphrase questions generated by rules or generation models. Jia and Liang (2017); Ribeiro et al. (2018a); Feng et al. (2018) focus on evaluating the over-stability issue by adding distracting sentences to the documents or reducing question word sequences.

However, the altered questions and documents are not natural texts and rarely appear in the real-world applications. In this paper, we collect the natural questions and documents to present the robustness challenge when applying neural models to the real-world applications. Specifically, (1) To evaluate the over-sensitivity issue, we leverage a retrieval-based approach to obtain paraphrase questions, that are true natural questions issued by people in Baidu Search. (2) To evaluate the over-stability issue, we ask people to annotate natural questions that have many words in common with a distracting sentence in the context. (3) To evaluate the generalization issue, we collect a test set that contains the documents from search results with the domain of K12 education and financial reports, while the training set consists of documents with general domain.

We further conduct extensive experiments based on DuReader$_{\text{robust}}$. The experimental results show that the models based on pre-trained LMs perform much worse than human does on the robustness test set, although they perform as well as human

on in-domain test set. Additionally, we have the following major findings about the behavior of existing models: (1) if a paraphrased question has more words rephrased from the original question, it is more likely that an MRC model gives a different answer. (2) if a question has more words appearing in a misleading sentence, it is more likely that an MRC model fails on the question. (3) the domain knowledge is a key factor that affects the generalization ability of MRC models.

In this paper, we have the following contributions:

- We create a dataset namely DuReader$_{\text{robust}}$ to comprehensively evaluate the robustness issues of neural MRC models, including *over-sensitivity*, *over-sensitivity* and *generalization*.
- We collect the true natural questions and documents to present the robustness challenge when applying the neural models to the real-world applications.
- We conduct extensive experiments to evaluate the MRC models based on pre-trained LMs. The experimental results demonstrate the robustness issues of these models, which may give the insights for future model development.

2 Dureader$_{\text{robust}}$ Dataset

In this section, we introduce DuReader$_{\text{robust}}$ dataset. It is constructed based on the large-scale Chinese machine reading comprehension dataset Dureader (He et al., 2017).

2.1 Training Set, Dev Set and In-Domain Test Set

In the original DuReader dataset (He et al., 2017), questions are real questions issued by users in Baidu Search, and documents are extracted from the search results of Baidu Search and Baidu Zhidao[1]. DuReader provides rich annotations for question types, including entity questions, description questions and yes/no questions. For each question, it provides as context multiple candidate documents containing multiple paragraphs. Formally, an instance in DuReader can be viewed as a quadruple of $\langle q, t, D, A\rangle$, where q is a question, t is its question type, D are its candidate documents (i.e. the context) and A are the reference answers.

[1]Baidu Zhidao is a web site of community question answering

For simplicity, in the construction of DuReader$_{\text{robust}}$, we reduce the context from multiple documents to one paragraph, and select the instances with the question type defined as entity. Thus, DuReader$_{\text{robust}}$ dataset can be considered as a set of triplets of $\langle q, p, A\rangle$, where q represents a question, p represents the paragraph that contains the reference answers A. To ensure the data quality, we remove incorrect triplets by employing crowdworkers to annotate all instances in the dataset, since the triplets automatically constructed from DuReader may lack enough context. In this way, we obtain a training set, a development set and an in-domain test set for DuReader$_{\text{robust}}$. The training set and development set contain 15K instances and 1.4K instances, respectively. There are 1.3K instances in the in-domain test set.

We say that this test set is **in-domain test set**, since its distribution is the same as the training set. In the traditional paradigm, we use the in-domain test set to evaluate model accuracy. By contrast, we need to have robustness test sets to evaluate the robustness issues of MRC models.

2.2 Robustness Test Set

As we discuss in the previous section, we aim to comprehensively evaluate the three robustness issues of neural MRC models, including over-sensitivity issue, over-stability issue and generalization issue. Hence, we construct a robustness test set in DuReader$_{\text{robust}}$, that includes three subsets: **over-sensitivity test set**, **over-stability test set** and **generalization test set**. In the following sections, we will illustrate the construction procedure in detail.

2.2.1 Over-sensitivity Test Set

Intuitively, if two questions are paraphrase of each other, a robust MRC model should give the same prediction. Otherwise, the model is said to be over-sensitive. Hence, we create an over-sensitivity test set to evaluate if a model has over-sensitivity issue. Previous work focuses on evaluating the over-sensitivity issue via question paraphrases, which are generated by rules or controlled generation models (Ribeiro et al., 2018b; Iyyer et al., 2018; Gan and Ng, 2019). However, the generated paraphrase questions might be different from the natural questions posed by people in the real-world applications, and their diversity might be limited.

Instead, we obtain paraphrase questions, which

are true natural questions issued by people in Baidu Search. We consider that this may help better evaluate the over-sensitivity of MRC models in real-world applications. Specifically, we use a retrieval-based toolkit that has two modules: (1) **paraphrase candidate retrieval.** Given a question, it will retrieve similar questions as candidates from an inverted index of search logs in Baidu Search. (2) **paraphrase similarity model**: Given a question and a retrieved paraphrase candidate, it will estimate the semantic similarity between them and determines if they are paraphrases of each other. The similarity model is a fine-tuned ERNIE (Sun et al., 2019) model by using a set of manually labeled paraphrase questions. This toolkit benefits from both the real search logs and a well-tuned similarity model. It can provide high-quality and diverse natural question paraphrases issued by people in Baidu Search[2]. Next, we will describe the procedure of collecting over-sensitivity test set.

First, we sample a set of instances $\{\langle q, p, A\rangle\}$ from the in-domain test set of DuReader$_{\text{robust}}$. For each q, we then obtain N question paraphrases $\{q'_1, q'_2, ..., q'_N\}$ using the toolkit. We further employ crowdworkers to remove all false paraphrases to ensure the data quality. We then replace q with the remaining paraphrased questions q'_i, and keep its paragraph p and the answer A unchanged. This will lead to new instances $\{\langle q'_i, p, A\rangle\}$. Finally, we select new instances to construct over-sensitivity test set, when the confidence scores of paraphrased questions given by the toolkit are higher than a threshold. This is a model independent way to collect data. Besides, we also employ a model dependent way. Specifically, we randomly choose one of MRC models used in Section 3 and use a new instance to attack the chosen model. If the model gives a different prediction from the original one, we select this instance to construct over-sensitivity test set.

The final over-sensitivity test set consists of both model-independent and model-dependent instances. There are 1.2K instances in total. The number of model-independent instances is equal to the number of model-dependent instances. Table 1a shows an instance in over-sensitivity test set.

[2]This toolkit is used internally at Baidu. We manually evaluate the question paraphrases given by the toolkit, and the accuracy is around 98%.

Algorithm 1: Annotate an instance for over-stability test set

Input: $\{\langle q, p, A\rangle\}$ tuple
Output: $\{\langle q^{'}, p, A\rangle\}$ tuple or null
Identify the named entities $\{e_1, ..., e_n\}$ together their entity types in p
Keep the named entities $\{e_i, ..., e_m\}$ that have the same types as A
if $0 < m < k$ **then**
  if *human experts consider the passage* p *contains a distracting sentence* **then**
    annotate the question q'
    return $\{\langle q', p, A\rangle\}$
  else return null;
else return null;

2.2.2 Over-stability Test Set

The complementary problem of *over-sensitivity* is *over-stability*: an over-stable model always relies on spurious lexical patterns without language understanding, and it over stably retains the predictions whenever the local contexts match the lexical patterns. Hence, the model is bad at distinguishing a sentence that answers the question from a distracting sentence that has many words in common with the question.

Previous work focuses on evaluating the *over-stability* issue by adding distracting sentences to the contexts or reducing question word sequences (Jia and Liang, 2017; Ribeiro et al., 2018a; Feng et al., 2018; Welbl et al., 2019). However, the altered questions and contexts are not natural texts and rarely appear in the real-world applications. In contrast to prior work, we ask people to annotate natural questions that have many words in common with a distracting sentence in the context. It better reflects the *over-stability* challenge that we have in the real-world applications.

First, we randomly sample a set of instances $\{\langle q, p, A\rangle\}$ from DuReader. Intuitively, a distracting sentence contains named entities that have the same types as A, since over-stable models rely on spurious patterns that usually match the correct answer types. Hence, we use a named entity recognizer[3] to identify all named entities in p together with their entity types. We keep the instances if there are named entities that have the same types as A. Then, we ask experts to annotate a new question q', if the experts consider that there are distracting sentences in p. The annotated question has much lexical overlap with a distracting sentence that does

[3]https://ai.baidu.com/tech/nlp_basic/lexical

not contain A in p. We say that $\{\langle q', p, A\rangle\}$ is a new instance. For each new instance, we randomly choose one of MRC models used in Section 3 and use the instance to attack the chosen model. If the model is failed, we will use the new instance to construct over-stability test set. The detailed procedure is shown in Algorithm 1.

In this way, we obtain 0.8k instances in over-stability test set. Table 1b shows an instance in over-stability test set.

2.2.3 Generalization Test Set

Previously, MRC models are primarily evaluated on in-domain test sets, while it is challenging to develop models that generalize well to new test distributions. Inspired by Fisch et al. (2019), we construct a generalization test set which has a different distribution from the training set.

Specifically, our generalization test set contains the data from two vertical domains, e.g. K12 education and financial reports. Next, we will give the detailed procedure of the data construction on these two domains.

K12 education. Like DuReader (He et al., 2017), we first collect questions whose intents are K12 education and their clicked documents from Baidu Search. The questions and documents contain the topics about mathematics, physics, chemistry, language and literature course. We then ask crowdworkers to annotate answers. In this way, we obtain 1.2K instances about K12 education.

Financial reports. Following Fisch et al. (2019), we leverage a dataset that is designed for information extraction in finance domain for MRC. The topics of this set includes management changes and equity pledge.

The original dataset contains the full texts of the financial reports as documents and the structured data that is extracted from the texts. Then, we use templates to generate questions for each data field in the structured data. Finally, we use these constructed instances for MRC. Each instance contains (1) a question generated from a template for a data field, (2) an answer that is the value in the data field and (3) a document from which the value (i.e. answer) is extracted. In this way, we obtain 0.4K instances about financial reports.

In total, we collect 1.6K instances in generalization test set. Table 1c shows an example.

Answer Type	%	Examples
Date	24.7	15分钟 (15 minutes)
Number	17.5	53.28厘米 (53.28cm)
Interval	11.8	1%至5% (1% to 5%)
Person	8.8	成龙 (Jackie Chan)
Organization	7.5	湖南卫视 (Hunan Satellite TV)
Money	7.0	2.7亿美元 (270 million dollars)
Location	6.0	巴西 (Brazil)
Software	2.2	百度地图 (Baidu Map)
Item	1.6	华为P9 (Huawei P9)
Other	12.9	群雄割据 (Heroic division)

Table 2: The frequency distribution of answer types in $\text{DuReader}_{\text{robust}}$.

2.3 Data Statistics

In a summary, $\text{Dureader}_{\text{robust}}$ consists of a training set, a development set, an in-domain test test and three robustness test sets. The data set contains 22K instances in total. The statistics of the dataset has been shown in Table 3. Additionally, the dataset covers a wide range of answer types (e.g. date, numbers, person). The frequency distribution of fine-grained answer types is shown in Table 2.

3 Experiments

In this section, we introduce the baselines based on pre-trained LMs, and conduct the experiments to examine the performance of these baselines on $\text{DuReader}_{\text{robust}}$. We further conduct extensive experimental analysis to obtain the insights about the robustness issues of these models.

3.1 Evaluation Metrics

To evaluate the held-out accuracy of an MRC model, we use the following two metrics, exact match (EM) and F1-score. Besides, we introduce different prediction ratio (DPR) as a metric to evaluate *over-sensitivity* of an MRC model. To calculate these metrics, we first normalize the predicted and reference answers by removing spaces and punctuation. We then do the calculation in Chinese character-level.

Exact match (EM). This metric measures the percentage of predicted answers that match any one of the reference answers exactly.

F1-score. (Rajpurkar et al., 2016) define this metric to measure the average overlap between the prediction and reference answer. (Rajpurkar et al., 2016) treat the prediction and reference as bags of tokens, and compute their F1. Instead, following Cui et al. (2018), we obtain the longest

Dataset	Paragraph len.	Question len.	Answer len.	# of instances
Train Set	291.88	9.19	5.39	14,520
Dev Set	288.16	9.38	6.66	1,417
In-domain Test Set	285.36	9.41	6.55	1,285
Robustness Test Set	132.09	11.97	7.33	3,556
All				**20,778**

Table 3: The statistics of $\text{DuReader}_{\text{robust}}$.

Models	H	L	A	# of Parameters	URL
$\text{BERT}_{\text{base}}$	768	12	12	110M	https://github.com/google-research/bert
ERNIE 1.0_{base}	768	12	12	110M	https://github.com/PaddlePaddle/ERNIE
ERNIE 2.0_{base}	768	12	12	110M	n/a
ERNIE 2.0_{large}	1024	24	16	340M	n/a

Table 4: The hyper-parameters of pre-trained language models. We denote the number of layers as L, the hidden size as H, and the number of self-attention heads as A. Besides, we provide the URL to download these pre-trained models.

common sequence (LCS) between them as their overlap and then compute the F1-score accordingly. We take the maximum F1 over all the reference answers for a given question, and then average over all the questions.

Different prediction ratio (DPR). This metric measures the percentage of paraphrase questions whose answers are different from the original questions. Formally, we define DPR of a neural MRC model $f(\theta)$ on a dataset D as follows.

$$DPR_D(f(\theta)) = \frac{\sum_{(q,q\prime)\in Q} \mathbb{1}[f(\theta;q) = f(\theta;q\prime)]}{\|Q\|},$$

where Q represents a set of pairs of an original question q and its paraphrase question $q\prime$ in dataset D. A high DPR score means the MRC model is sensitive with respect to question paraphrases.

3.2 The Experimental Settings

Baselines. In this paper, we have four baselines that are based on pre-trained LMs, including $\text{BERT}_{\text{base}}$, ERNIE 1.0_{base}, ERNIE 2.0_{base} and ERNIE 2.0_{large}. The hyper-parameters of these pre-trained models have been listed in Table 4.

Hyper-parameters. In the fine-tuning stage, we use the same hyper-parameters for all models. The learning rate is 3e-5, and batch size is 32. We set the number of epochs is 5. The maximal answer length and the maximal document length is 20 and 512, respectively. We set length of document stride is 128.

Human Performance. We evaluate human performance on both the in-domain test set and robustness test set. Specifically, we sample two hundred examples from in-domain test set and one hundred examples from each of over-sensitivity, over-stability and generalization set. We ask crowdworkers to annotate answers to these sampled instances. Then, we calculate the EM and F1-scores of these annotated examples as the human performance.

3.3 The Main Results

The performance of the baseline systems on the dev set, in-domain test set and robustness test set are shown in Table 5. We can observe that the performance of these baselines is approaching human on in-domain test set, while the gap between baseline performance and human performance on robustness test set is much larger. The performance of these baselines on robustness test set is much lower than their performance on in-domain test set (although they are not fully comparable). In contrast, the human performance is closed on two test sets. Besides, ERNIE 2.0_{large} performs the best on both in-domain test set and robustness test set, while $\text{BERT}_{\text{base}}$ and ERNIE 1.0_{base} perform the worst.

We further analyze the performance of the baselines on the three subsets of robustness test set. The results have been shown in Table 6. We can observe that the baseline performance significantly drops on the over-stability test set and the generalization test set. In contrast, the baseline performance on the over-sensitivity test set does not drop too much. However, the different prediction ratios of baselines are large as we will show in Section 3.4.1.

3.4 Experimental Analysis

In this section, we give more detailed analysis of the baseline behavior on the three subsets of robustness test set.

	In-domain dev set		In-domain test set		Robustness test set	
	EM	**F1**	**EM**	**F1**	**EM**	**F1**
BERT$_{\text{base}}$	71.20	82.87	67.70	80.85	37.57	53.86
ERNIE 1.0$_{\text{base}}$	68.73	81.12	66.72	80.50	36.75	55.64
ERNIE 2.0$_{\text{base}}$	70.23	81.69	67.23	81.10	38.89	58.42
ERNIE 2.0$_{\text{large}}$	72.74	84.68	68.87	82.45	43.16	60.92
Human			78.00	89.75	72.00	86.43

Table 5: The experimental results on in-domain dev set, in-domain test set and robustness test set.

	Over-sensitivity test set		Over-stability test set		Generalization test set	
	EM	**F1**	**EM**	**F1**	**EM**	**F1**
BERT$_{\text{base}}$	53.31	69.30	16.78	38.40	36.41	50.15
ERNIE 1.0$_{\text{base}}$	58.10	73.89	17.27	38.34	32.86	52.84
ERNIE 2.0$_{\text{base}}$	56.75	73.90	23.65	43.88	33.18	54.17
ERNIE 2.0$_{\text{large}}$	57.93	74.31	29.04	47.69	39.25	57.59

Table 6: The experimental result three robustness test subsets.

	DPR (%)
BERT$_{\text{base}}$	22.73
ERNIE 1.0$_{\text{base}}$	19.88
ERNIE 2.0$_{\text{base}}$	19.54
ERNIE 2.0$_{\text{large}}$	16.52

Table 7: The different prediction ratio of baseline systems on over-sensitivity test set.

3.4.1 Over-sensitivity Analysis

In this section, we first measure DPRs of baselines on paraphrased questions. Table 7 gives the results on over-sensitivity test set. We can observe that the four baselines give around 16% to 22% different predictions. This means that these baselines are sensitive to the paraphrased questions. The large model (e.g. ERNIE 2.0$_{\text{large}}$ shows less sensitivity than the base models.

Next, we try to figure out what kind of paraphrases lead to different predictions. We first define five types of question paraphrases: (1) **Word reordering.** Reorder words in the original question to a new sequence. (2) **Function words.** Change function words in the original questions. (3) **Synonym.** Use synonym to substitute words in the original question. (4) **Content words.** Add or remove content words in the original questions. (5) **Complex.** There are more than one previously defined types of changes. We then randomly sample one hundred instances and obtain the predictions by ERNIE 2.0$_{\text{large}}$. As shown in Table 8, we can observe that most of changed predictions come from **synonym**, **content words** and **complex**. This analysis suggests that the models are sensitive to the changes of content words.

We further conduct analysis to examine a hypothesis, that if there are more words changed in the questions, it is more likely that the predicted answers will be changed. We use f1-score to measure the similarity between paraphrased questions and original questions. If f1-score is low, it means that many words in the original question have been changed. We split the paraphrased questions into buckets according to their similarity to the original questions, and then we see if there is any relation between DPR and f1-score similarity. The experimental results have been shown in Figure 1. Overall, we can observe that DPRs of all the baselines are negatively correlated to the F1-score similarity between original questions and paraphrased questions. The results verify the hypothesis.

From Table 6, we also observe that the accuracy of baselines (in terms of EM and F1) drops on the over-sensitivity test set. We want to know why the accuracy has decreased on the over-sensitivity test set. If we treat both questions and documents as bags of words, the documents can recall 74.73% words in the original questions, and 70.95% words in the paraphrased questions, respectively. It means that there is less overlap between paraphrased questions and documents. The mismatch between paraphrased questions and documents brings the decrease on accuracy.

3.4.2 Over-stability Analysis

In previous section, we discuss that MRC models can be easily mislead by distracting sentences. Given a document and a question, a distracting sentence is a sentence containing no ground-truth answer in the document, yet it has the highest F1-score similarity to the question. Besides, it contains

Types	# of changes (%)	# of same (%)	Total #
Word reordering	1 (12.50)	7 (87.50)	8
Function words	0 (00.00)	4 (100.00)	4
Synonym	7 (20.00)	28 (80.00)	35
Content words	5 (16.00)	25 (52.94)	30
Complex	8 (34.78)	15 (65.21)	23

Table 8: The relation between paraphrase types and the prediction changes.

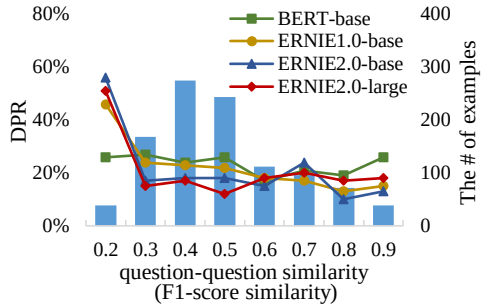

Figure 1: The correlation between DPR and question-question similarity (F1-score similarity) on over-sensitivity test set.

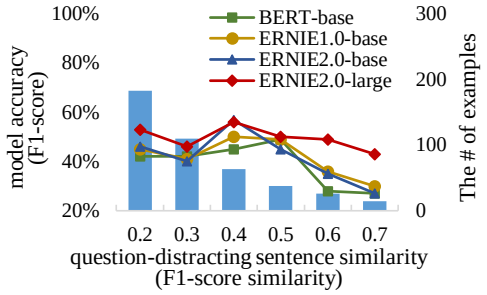

Figure 2: The correlation between the model accuracy (F1-score) and question-distracting sentence similarity (F1-score similarity) on over-stability test set.

named entities with the same type as the ground-truth answer.

Intuitively, if the similarity between a distracting sentence and a question is high, it is likely that a model will be mislead by the distracting sentence. We conduct the analysis to examine this hypothesis. Specifically, we split the paraphrased questions in to buckets according to their similarity to the distracting sentences, and then we see if there is any relation between model accuracy (in terms of F1-score) and question-distracting sentence similarity (in terms of F1-score similarity). Figure 2 shows the analysis results. We can observe that when the similarity between questions and distracting sentences becomes higher, the model accuracy (in terms of F1-score) becomes lower. Additionally, we observe that the large model (ERNIE 2.0$_{\text{large}}$) is less over-stable than the small models. In other words, the small models are likely being mislead by lexical pattern matching.

3.4.3 Generalization Analysis

Overall, the performance of baselines on generalization test set is lower than their performance on the in-domain test set. The generalization test set consists of the data from two domains: financial reports and K12 education. Table 9 shows the performance of baselines on the two domains separately. We observe that the baselines perform similarly on these two domains. We further conduct detailed analysis of model behavior on these two domains, respectively.

K12 Education. We divide the data of education into four topics including math, chemistry, language and others. Table 10 shows the performance of ERNIE 2.0 $_{\text{large}}$ on these four topics. We can observe that the model performs the worst on math and chemistry. The model cannot generalize well on these two topics, since the training set does not contain the knowledge about math equation or chemical equation. In contrast, the model performs better on language and others. Because the model learns relevant knowledge and patterns on training set.

Financial Reports. The data of financial reports contains management changes and equity pledge. The performance of ERNIE 2.0$_{\text{large}}$ on management changes and equity pledge is 68.63% and 49.15% respectively. The model generalizes well on management changes, since the training set contains the relevant knowledge and patterns about asking person names. In contrast, the model performs worse on equity pledge. We classify the data of equity pledge into five sets according to the question types. Table 11 shows the performance of ERNIE 2.0 $_{\text{large}}$ on the five question types. We can observe that the model performs the worst on the questions about company abbreviations, pledgee and pledgor, since there is little domain knowledge in the training set. In contrast, the model performs better on the questions about amount and date, since the model already learns relevant patterns and knowledge in

	Financial Reports		Education	
	EM	F1	EM	F1
BERT$_{base}$	30.73	51.16	38.70	50.83
ERNIE 1.0$_{base}$	26.53	50.53	34.67	53.11
ERNIE 2.0$_{base}$	24.30	53.30	35.26	54.15
ERNIE 2.0$_{large}$	38.26	56.93	39.54	57.79

Table 9: The performance of educational and financial fields.

Topcis	EM	F1	#
Math	19.85	31.71	136
Chemistry	27.86	50.84	323
Language	38.82	56.39	255
Others	57.76	73.76	438
All	40.71	58.52	1152

Table 10: The performance of ERNIE 2.0$_{large}$ on the four topics in the domain of education.

Question Types	EM	F1	#
Company abbreviations	0	17.12	18
Pledgee	15.38	43.77	26
Pledgor	8.00	20.56	25
The pledge amount	17.34	51.97	98
Others (e.g. pledge date)	60.41	73.19	48
All	24.18	49.15	215

Table 11: The performance of ERNIE 2.0$_{large}$ on the five topics in the domain of financial reports.

the training set.

In a summary, we can see that domain knowledge is a key factor that affects the generalization ability of MRC models.

4 Related Work

4.1 Machine Reading Comprehension Datasets

With the increasing availability of large scale MRC datasets, there has been great advancements in MRC techniques. SQuAD 1.0 (Rajpurkar et al., 2016) is the first large-scale MRC dataset consisting of questions and answers annotated by crowdworkers. It is designed for extractive MRC that requires the machine to locate the correct answer span to a question in a given context document.

Then, a number of MRC datasets have been created to challenge MRC models from different aspects: (1) Rajpurkar et al. (2018) proposes SQuAD 2.0, that consists of both answerable and unanswerable questions and requires the machine to determine when no answer is supported by the context. (2) The datasets of MSMARCO (Nguyen et al., 2016), DuReader (He et al., 2017) and Natural Questions (Kwiatkowski et al., 2019) require the machine to extract answers from multiple passages in the scenario of search. (3) CoQA (Reddy et al., 2019) and QuAC (Choi et al., 2018) are designed for conversational MRC that consists of conversational questions on a set of articles. (4) HotpotQA (Yang et al., 2018) and Qangaroo (Welbl et al., 2018) require multi-hop reasoning. DROP (Dua et al., 2019) requires discrete reasoning over paragraphs of text. (5) CommonsenseQA (Talmor et al., 2018), ReCoRD (Zhang et al., 2018) and ARC (Clark et al., 2018) require commonsense or external knowledge for machine reading.

The above mentioned datasets try to evaluate the reading comprehension ability of machine from different aspects, while this paper focus on creating a dataset that evaluating the robustness of neural MRC models.

4.2 Over-sensitivity

The *over-sensitivity* means that semantically equivalent paraphrase questions can alter the predictions of a model, when it should not. The *over-sensitivity* issue of MRC models has been studied recently. Previous work mainly focuses on evaluating the *over-sensitivity* issue via question paraphrases, that are generated by rules or controlled generation models. Ribeiro et al. (2018b) propose mining high quality semantically equivalent adversarial rules to generate question paraphrases by involving human-in-the-loop. Iyyer et al. (2018) propose a syntactically controlled paraphrase networks to generate paraphrase as adversarial examples. Gan and Ng (2019) propose a method to generate diverse paraphrased questions by guiding the generation model with paraphrase suggestions.

However, the generated paraphrase questions might be different from the natural questions posed by people in the real-world applications. Instead, we use a retrieval-based approach to acquire paraphrase questions, that are true natural questions issued by people in Baidu Search. We consider that this may help better evaluate the over-sensitivity of MRC models in real-world applications.

4.3 Over-stability

The complementary problem of *over-sensitivity* is *over-stability*: an over-stable MRC model relies on spurious patterns too much without language understanding, and it locates a false answer in a distracting sentence that matches the patterns.

Previous work focuses on evaluating the *over-stability* issue by adding distracting sentences to the contexts or reducing question word sequences. Jia and Liang (2017) propose adding a distracting sentence to the original context, so as to confuse models. Ribeiro et al. (2018a) develop a toolkit to identify minimal feature sets in questions that keep predictions always the same. Feng et al. (2018) propose a gradient-based method to reduce questions to minimal word sequences without changing predictions of models. Welbl et al. (2019) searches among semantic variations of the question for which a model erroneously predicts the same answer.

However, the altered questions and contexts are not natural and rarely appear in the real-world applications. In contrast to prior work, we ask people to annotate alternative questions that have many words in common with a distracting sentence in the context. It better reflects the *over-stability* challenge that we have in the real-world applications.

4.4 Generalization

Previously, MRC models are primarily evaluated on in-domain test sets. It is challenging to develop models that generalize well to new test distributions. Fisch et al. (2019) propose a shared task that focuses on evaluating generalization of MRC models. Specifically, it requires that the models are trained on a training set pooled from six MRC datasets, and are evaluated on other twelve different test datasets.

Inspired by the previous work, we collect a test set that contains the documents with the domain of K12 education and financial reports from Baidu Search, while the training set consists of documents with general domain. Our experimental results show that domain knowledge is a key factor that affects the generalization ability of MRC models.

4.4.1 Summary

As a short summary, we focus on creating a dataset to comprehensively evaluate the robustness of MRC models, including *over-sensitivity*, *over-stability* and *generalization*. Comparing to previous work, we collect the natural questions and documents to evaluate the robustness challenge of neural models when applying them to the real-world applications.

5 Conclusion

In this paper, we introduce a Chinese data set, namely $\text{DuReader}_{\text{robust}}$ to comprehensively evaluate the robustness of MRC models when applying them to the real-world applications. Specifically, we focus on challenging MRC models from the following aspects: (1) over-sensitivity, (2) over-stability, and (3) generalization. The previous work studies these problems mainly by altering the inputs to unnatural texts. By contrast, the advantage of $\text{DuReader}_{\text{robust}}$ is that its questions and documents are natural texts from real-world applications. This presents the true robustness challenges in the real-world applications. Our experiments show that the MRC models based on the pre-trained language models perform poorly on the robustness test set, while they perform well on the in-domain test set. We also conduct extensive experiments to examine the behavior of existing models on the robustness test set.

References

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018. A span-extraction dataset for chinese machine reading comprehension.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *NAACL-HLT*.

Shi Feng, Eric Wallace, Alvin Grissom, Mohit Iyyer, Pedro Rodriguez, and Jordan L. Boyd-Graber. 2018. Pathologies of neural models make interpretation difficult. In *EMNLP*.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. Mrqa 2019 shared task: Evaluating generalization in reading comprehension.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *ACL*.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2017. Dureader: a chinese machine reading comprehension dataset from real-world applications.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *NAACL-HLT*.

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. volume 7, pages 453–466. MIT Press.

Zhen-Zhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. volume abs/1909.11942.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. volume abs/1907.11692.

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.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. volume 7, pages 249–266. MIT Press.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018a. Anchors: High-precision model-agnostic explanations. In *AAAI*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018b. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge.

Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer.

Johannes Welbl, Pasquale Minervini, Max Bartolo, Pontus Stenetorp, and Sebastian Riedel. 2019. Undersensitivity in neural reading comprehension. volume abs/2003.04808.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. volume 6, pages 287–302. MIT Press.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. Record: Bridging the gap between human and machine commonsense reading comprehension.