DuReader_{robust}: A Chinese Dataset Towards Evaluating Robustness and Generalization of Machine Reading Comprehension in Real-World Applications

Hongxuan Tang[1]* **Hongyu Li**[2] **Jing Liu**[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 task in natural language processing and has achieved remarkable advancements. However, most of the neural MRC models are still far from robust and fail to generalize well in real-world applications. In order to comprehensively verify the robustness and generalization of MRC models, we introduce a real-world Chinese dataset – DuReader_{robust}. It is designed to evaluate the MRC models from three aspects: over-sensitivity, over-stability and generalization. Comparing to previous work, the instances in DuReader_{robust} are natural texts, rather than the altered unnatural texts. It presents the challenges when applying MRC models to real-world applications. The experimental results show that MRC models do not perform well on the challenge test set. Moreover, we analyze the behavior of existing models on the challenge test set, which may provide suggestions for future model development. The dataset and codes are publicly available at https://github.com/baidu/DuReader.

1 Introduction

Machine reading comprehension (MRC) requires machines to comprehend text and answer questions about it. With the development of deep learning, the recent studies of MRC have achieved remarkable advancements (Seo et al., 2017; Wang and Jiang, 2017; Devlin et al., 2019; Liu et al., 2019; Lan et al., 2020). However, previous studies show that most of the neural models are not robust enough (Jia and Liang, 2017; Ribeiro et al., 2018b; Talmor and Berant, 2019a; Welbl et al., 2020) and fail to generalize well (Talmor and Berant, 2019b).

---

* This work was done while the first author was doing internship at Baidu Inc.
† Corresponding authors

To further promote the studies of robust and well generalized MRC, we construct a Chinese dataset – DuReader_{robust} which comprises natural questions and documents. In this paper, we focus on evaluating the robustness and generalization from the following aspects, where robustness consists of over-sensitivity and over-stability:

(1) **Over-sensitivity** denotes that MRC models provide different answers to the paraphrased questions. It means that the models are overly sensitive to the difference between the original question and its paraphrased question. We provide an example in Table 1a.

(2) **Over-stability** means that the models might fail into a trap span that has many words in common with the question, and extract an incorrect answer from the trap span. Because the models overly rely on spurious lexical patterns without language understanding. We provide an example in Table 1b.

(3) **Generalization**. The well-generalized MRC models have good performance on both in-domain and out-of-domain data. Otherwise, they are less generalized. We provide an example in Table 1c.

In previous work, the above issues have been studied separately. In this paper, we aim to create a dataset namely DuReader_{robust} to comprehensively evaluate the three issues of neural MRC models. Previous work mainly studies these issues by altering the questions or the documents. Ribeiro et al. (2018b); Iyyer et al. (2018); Gan and Ng (2019) evaluate the over-sensitivity issue via paraphrase questions generated by rules or generative models. Jia and Liang (2017); Ribeiro et al. (2018a); Feng et al. (2018); Talmor and Berant (2019a) 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. It is not clear that how the evaluation based

Passage	Passage
近年来，随着琥珀蜜蜡市场的兴起，蜜蜡与琥珀的价格都有不断上涨的趋势，其中蜜蜡首饰的价格一般是琥珀首饰价格的2–4倍，最近几年二者价格差距更大…….	*In recent years, with the rise of the amber market, the price of amber keeps going up. The price of opaque amber is generally 2–4 times the price of clear amber ...*
Original Question	**Original Question**
琥珀和蜜蜡哪一个比较贵	*Which is more expensive, clear amber or opaque amber?*
Golden Answer：蜜蜡	**Golden Answer** : *opaque amber*
Predicted Answer：蜜蜡 (BERT_{base})	**Predicted Answer** : *opaque amber* (BERT$_{base}$)
Paraphrase Question	**Paraphrase Question**
蜜蜡和琥珀哪个价格高	*Which has the higher price, opaque amber or clear amber?*
Golden Answer：蜜蜡	**Golden Answer** : *opaque amber*
Predicted Answer：琥珀 (BERT$_{base}$)	**Predicted Answer** : *clear amber* (BERT$_{base}$)

(a) An example illustrates the over-sensitivity issue, where BERT$_{base}$ gives different predictions to the original question and the paraphrased question.

Passage	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。	*Many people argue that the zongzi (rice dumpling) leaves are made of hemp. Actually, it is the palm tree, the real origin, that endows zongzi with the special pronunciation.*
Question	**Question**
包粽子的线来自什么	*What is the raw material of zongzi leaves?*
Golden Answer：棕榈树	**Golden Answer** : *palm tree*
Predicted Answer：麻叶子 (BERT$_{base}$)	**predicted Answer** : *hemp* (BERT$_{base}$)

(b) An example illustrates the over-stability issue. The underlined span in the passage appears as a trap because it has many words in common with the question. BERT$_{base}$ falls into the trap.

Passage	Passage
cos(2x)'=-sin(2x)(2x)'=-2sin(2x)* 属于复合函数的求导。	*cos(2x)'=-sin(2x)*(2x)'=-2sin(2x) This is the derivative of a compound function.*
Question	**Question**
*cos2x*的导数是多少？	*What is the derivative of cos2x?*
Golden Answer : *-2sin(2x)*	**Golden Answer** : *-2sin(2x)*
Predicted Answer : *-sin(2x)* (BERT$_{base}$)	**Predicted Answer** : *-sin(2x)* (BERT$_{base}$)

(c) An example illustrates the generalization issue. Although BERT$_{base}$ is sufficiently trained on large-scale open-domain data, it fails to predict the answer to a math question.

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

on such unnatural texts can help the improvements of the neural models in real-world applications. By contrast, all the instances in DuReader$_{robust}$ are natural texts and collected from the Baidu search.

We conduct extensive experiments based on DuReader$_{robust}$. The experimental results show that the models based on pre-trained language models (LMs) (Devlin et al., 2019; Sun et al., 2019; Liu et al., 2019) do not perform well on the challenge set. Besides, we have the following findings on the behaviors of the models: (1) if a paraphrased question contains more words rephrased from the original question, it is more likely that MRC models provide different answers; (2) the trap spans which share more words with the questions easily mislead MRC models; (3) domain knowledge is a key factor that affects the generalization ability of MRC models.

2 Dataset: DuReader$_{robust}$

DuReader$_{robust}$ is built on DuReader, a large-scale Chinese MRC dataset (He et al., 2018). In DuReader, all questions are issued by real users of

Dataset	len(p)	len(q)	len(a)	#
Train	291.88	9.19	5.39	14,520
Development	288.16	9.38	6.66	1,417
Test	285.36	9.41	6.55	1,285
Challenge	132.09	11.97	7.33	3,556
All				20,778

Table 2: The statistics of DuReader$_{robust}$.

Baidu search, and the document-level contexts are collected from search results. In DuReader$_{robust}$，we select entity questions and paragraph-level contexts from DuReader. We further employ crowd-workers to annotate the answer span conditioned on the question and the paragraph-level context [1]. Additionally, we used a mechanism to ensure data quality, where 10% of the annotated data will be randomly selected and reviewed by linguistic experts. If the accuracy is lower than 95%, the crowd-workers need to revise all the answers until the accuracy for the randomly selected data is higher than 95%.

---
[1]The instances which have insufficient contexts for answering the questions are discarded.

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	北京 (Beijing)
Software	2.2	百度地图 (Baidu Map)
Item	1.6	华为P9 (Huawei P9)
Other	12.9	管理学 (Management Science)

Table 3: The frequency distribution and examples of different answer types in DuReader$_{\text{robust}}$.

Eventually, we collect about 21K instances for DuReader$_{\text{robust}}$, each of which is a tuple $\langle q, p, A \rangle$, where q is a question, p is a paragraph-level context containing reference answers A. Similar to the existing MRC datasets, DuReader$_{\text{robust}}$ consists of training set, in-domain development set and in-domain test set, whose sizes are 15K, 1.4K and 1.3K respectively. Besides, DuReader$_{\text{robust}}$ contains a challenge test set, in which 3.5K instances are created to evaluate the robustness and generalization of MRC models. The challenge test set can be divided into three subsets including over-sensitivity set, over-stability set and generalization set. Table 2 shows the statistics of DuReader$_{\text{robust}}$. Besides, DuReader$_{\text{robust}}$ covers a wide range of answer types (e.g. date, numbers, person, etc.). The frequency distribution and examples of the answer types are shown in Table 3. Next, we will present our way to construct the three subsets in the challenge test set.

2.1 Over-sensitivity Subset

We build the over-sensitivity subset in the following way. First, we sample a subset of instances $\{\langle q, p, A \rangle\}$ from the in-domain test set of DuReader$_{\text{robust}}$. For each question q, we obtain its N paraphrases $\{q'_1, q'_2, ..., q'_N\}$ using the paraphrase retrieval toolkit (See Appendix A for further details). To ensure the paraphrase quality, we employ crowd-workers to discard all false paraphrases. Then, we replace q with the paraphrased question q'_i, and keep the original context p and answers A unchanged. This leads to the new instances $\{\langle q'_i, p, A \rangle\}$, and they are used as the model-independent instances in the over-sensitivity subset. Besides, we also employ a model-dependent way to collect instances. Specifically, we use paraphrased instances to attack the MRC models based on ERNIE (Sun et al., 2019) and RoBERTa (Liu

Algorithm 1: Annotate an instance for over-stability subset

Input: $\{\langle q, p, A \rangle\}$ tuple
Output: $\{\langle q', p, A' \rangle\}$ tuple or null
Identify the named entities $\{e_1, ..., e_n\}$ along with their entity types in p
Keep the named entities $\{e_i, ..., e_m\}$ with the same types as A
if $1 < m < k$ **then**
  if *linguistic experts consider the passage p contains a trap* **then**
    annotate a new question q' and answers A'
    A and A' share the same named entity type
    return $\{\langle q', p, A' \rangle\}$
  else return null;
else return null;

et al., 2019). If one of the models gives a different prediction from the predicted answer of the original question, we adopt the instance, otherwise we discard it. The instances collected in the above model-dependent and model-independent ways constitute the over-sensitivity subset. The over-sensitivity subset consists of 1.2K instances. The number of model-independent instances is equal to that of model-dependent instances. Table 1a shows an example in the over-sensitivity subset.

2.2 Over-stability Subset

Intuitively, a trap span that has many words in common with the questions may easily mislead MRC models. Following this intuition, the over-stability subset is constructed as follows. First, we randomly select a set of instances $\langle q, p, A \rangle$ from DuReader. In general, a trap span may contain non-answer named entities of the same type as the reference answers A. This is because over-stable models usually rely on spurious patterns that match the correct answer types. Thus, we use a named entity recognizer [2] to identify all named entities in p along with their entity types. We keep the corresponding instance, if there are non-answer named entities that are of the same type as A. Then, we ask linguistic experts to annotate a new question q' and answers A', if they consider p contains trap spans. A and A' share the same named entity type. The annotated question q' has a high level of lexical overlap with a trap span that does not contain A. We say $\{\langle q', p, A' \rangle\}$ can be considered as a candidate instance. Each candidate instance is used to attack one of the MRC models based on ERNIE (Sun et al., 2019) and RoBERTa (Liu et al.,

---

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

	In-domain dev set		In-domain test set		Challenge test set	
	EM	F1	EM	F1	EM	F1
BERT_{base}	71.20	82.87	67.70	80.85	37.57	53.86
ERNIE 1.0_{base}	68.73	81.12	66.72	80.50	36.75	55.64
RoBERTa_{large}	74.17	86.02	71.20	84.16	45.02	62.83
Human			78.00	89.75	72.00	86.43

Table 4: Comparing MRC baselines to human on the development, test and all challenge sets.

	Over-Sensitivity		Over-Stability		Genera-lization	
	EM	F1	EM	F1	EM	F1
BERT_{base}	53.31	69.30	16.78	38.40	36.41	50.15
ERNIE 1.0_{base}	58.10	73.89	17.27	38.34	32.86	52.84
RoBERTa_{large}	55.24	75.16	28.18	47.03	46.03	61.67

Table 5: The results on the three subsets of the challenge set.

	DPR (%)
BERT_{base}	22.73
ERNIE 1.0_{base}	19.88
RoBERTa_{large}	16.44

Table 6: The DPRs of baselines on the over-sensitivity subset.

2019). The candidate instance will be used to construct an overstability subset, if one of the model fails. Algorithm 1 shows the detailed procedure (See Appendix B for details). As a result, we have 0.8K instances to evaluate over-stability. Table 1b shows an example from the over-stability subset.

2.3 Generalization Subset

The in-domain test set consists merely of in-domain data (i.e., the distribution is the same as the one in the training and development sets). In order to evaluate the generalization ability of MRC models, we construct a generalization subset which comprises out-of-domain data. The out-of-domain data is collected from two vertical domains. The details are as follows.

Education We collect educational questions and documents from Baidu search, and we ask crowdworkers to annotate 1.2K high-quality tuples $\langle q, p, A \rangle$. The topics include mathematics, physics, chemistry, language and literature. Table 1c shows an example.

Finance Following Fisch et al. (2019), we leverage a dataset that was originally designed for information extraction in the finance domain for MRC. We obtain 0.4K instances of financial reports this way. The construction details are presented in Appendix C.

3 Experiments

3.1 Baselines and Evaluation Metrics

We consider three baseline models in the experiments. They are based on different pre-trained language models, including BERT_{base} (Devlin

et al., 2019), ERNIE 1.0_{base} (Sun et al., 2019) and RoBERTa_{large} (Liu et al., 2019). In Appendix D, we set the hyperparameters of our baseline models.

Following Rajpurkar et al. (2016), we use exact match (EM) and F1-score to evaluate the held-out accuracy of an MRC model. All the metrics are calculated at Chinese character level, and we normalize both the predicted and true answers by removing spaces and punctuation marks.

3.2 Main Results

Table 4 shows the baseline results on the in-domain development set, in-domain test set, and challenge test set. The baseline performance is close to human performance on the in-domain test set, whereas the gap between baseline performance and human performance on the challenge test set is much larger. In Appendix E, we describe the method for calculating human performance.

We further evaluate the baselines on the three challenge subsets for over-sensitivity, over-stability and generalization separately. Table 5 shows the results. We have found that baseline performance declines significantly for over-stability and generalization subsets (compared to the "In-domain test set" in Table 4). In contrast, the baseline performance degrades less significantly on the over-sensitivity subset, although there is still a noticeable gap.

3.3 Discussion 1: Over-sensitivity

First, we calculate the different prediction ratios (DPRs) of the baselines on the over-sensitivity subset. DPR measures the percentage of the paraphrased questions that yield different predictions. DPR is formulated in Appendix F. Table 6 presents the DPRs of the baselines on the over-sensitivity subset. The baselines obtained around 16% to 22% DPRs, which demonstrates that the baselines are sensitive to part of the paraphrased questions.

Second, we examine a hypothesis - if a paraphrased question contains more words rephrased from the original question, the MRC model is more likely to produce different answers. To measure how similar paraphrased questions are to the original questions, we use the F1-score. A low F1-score
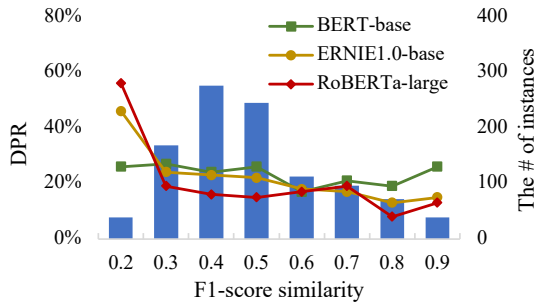
Figure 1: The correlation between DPR and F1-score based question similarity on the over-sensitivity subset.
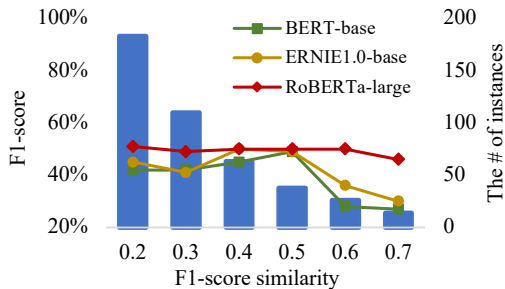

Figure 2: The correlation between the model performance and question-trap similarity on the over-stability subset.

means that many words in the original question have been rephrased. We divide the paraphrased questions into buckets based on how similar they are to the original questions, and we then examine whether there is correlation between DPR and F1-score similarity. Based on Figure 1, we can observe that the DPRs of all the baselines are negatively correlated with the F1-score similarity between the original and paraphrased questions. The results confirm the hypothesis.

3.4 Discussion 2: Over-stability

MRC models might be easily misled by trap spans that share many words with the questions. We examine whether there is a correlation between MRC performance (F1-score) and question-trap similarity in this section. Based on the similarity between trap spans and questions, we divide trap spans into buckets. According to Figure 2, the performance of the base models decreases as similarity increases and the large model (RoBERTa$_{large}$) is less over-stable than the base ones.

3.5 Discussion 3: Generalization

Table 7 shows the baseline performance in the the domains of finance and education. We can observe

	Finance		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
RoBERTa$_{large}$	40.22	61.16	47.77	61.82

Table 7: The performance of baselines in the domains of education and finance.

Topcis	EM	F1	#
Math	19.85	34.63	136
Chemistry	37.46	53.88	323
Language	44.31	61.18	255
Others	69.63	79.28	438
All	49.13	62.88	1152

Table 8: The performance of baselines on different topics in the domain of education.

that the baselines perform poorly for both domains. Additionally, we examine how baseline models behave in the education domain. Table 8 shows the performance of RoBERTa$_{large}$ on the four topics in the education domain. The model performs much worse when it comes to math and chemistry, since these topics are rare in the training set. The results of this analysis suggest that domain knowledge is a key factor affecting the generalization ability of MRC models. More discussion can be found in Appendix G.

4   Conclusion

In this paper, we create a Chinese dataset – DuReader$_{robust}$ and use it to evaluate both the robustness and generalization of the MRC models. Its questions and documents are natural texts from Baidu search. This presents the robustness and generalization challenges in the real-world applications. Our experiments show that the MRC models based on the pre-trained LMs do not perform well on DuReader$_{robust}$ challenge set. We also conduct extensive experiments to examine the behaviors of the MRC models on the dataset and provide insights for future model development.

Acknowledgement

Ethical Considerations

We aim to provide researchers and developers with a dataset DuReader$_{robust}$ to improve the robustness and generalization ability of MRC models. We also take the potential ethical issues into account. (1) All the instances in the DuReader$_{robust}$ have been desensitised. (2) Regarding to the issue of labor compensation, we make sure that all the crowdsourcing workers are fairly compensated.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong, China. Association for Computational Linguistics.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia. Association for Computational Linguistics.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. 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 1875–1885.

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 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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. In *arXiv preprint arXiv:1907.11692*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018a. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Marco Túlio 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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 856–865. Association for Computational Linguistics.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

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. *arXiv preprint arXiv:1904.09223*.

Alon Talmor and Jonathan Berant. 2019a. Multiqa: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4911–4921. Association for Computational Linguistics.

Alon Talmor and Jonathan Berant. 2019b. Multiqa: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4911–4921. Association for Computational Linguistics.

Shuohang Wang and Jing Jiang. 2017. Machine comprehension using match-lstm and answer pointer. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Johannes Welbl, Pasquale Minervini, Max Bartolo, Pontus Stenetorp, and Sebastian Riedel. 2020. Undersensitivity in neural reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 1152–1165. Association for Computational Linguistics.

A Paraphrase Retrieval Toolkit

We use a paraphrase retrieval toolkit to obtain paraphrase questions. The toolkit is used internally at Baidu, and our manual evaluations show that the accuracy of the retrieval results is around 98%. The paraphrase retrieval toolkit consists of two basic modules as follows.

- **Paraphrase Candidate Retriever** The retriever is a light-weight module. Given a question, the retriever will retrieve top-k paraphrase candidates from the search logs of Baidu Search.

- **Paraphrase Candidate Re-ranker** The re-ranker is a model fine-tuned from ERNIE (Sun et al., 2019) by using a set of manually labeled paraphrase questions. Given a set of retrieved paraphrase candidates, the re-ranker will estimate the semantic similarity between the original question and the paraphrase candidates. If the semantic similarity is higher than a pre-defined threshold, the candidate will be used as a paraphrased question.

B The Illustration of Annotating Over-stability Instances

Figure 3 illustrates the annotation of an overstability instance. In the instance, the answer to the original question is 30-40 minutes. The entity type of 5-10 minutes is the same as 30-40 minutes. The annotator raise a new question by revising the original question, the answer to the new question is 5-10 minutes. The sentence contains 30-40 minutes has many words in common with the new question, and it is considered as a trap sentence. The new question may mislead the model to predict the answer to the new question as 30-40 minutes.

C The Construction of Finance Data

We leverage a dataset that is originally designed for information extraction in finance domain. 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.
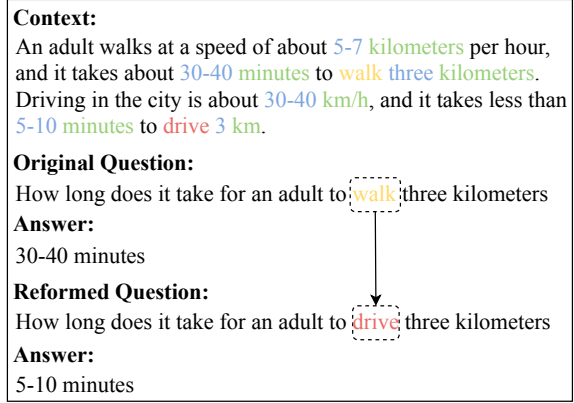

Figure 3: The illustration of annotating an overstability instance.

D Hyperparameters

We use a number of pre-trained language models in our baseline systems. When fine-tuning different pre-trained language models, we use the same hyperparameters. The settings of hyperparameters are as follows. The learning rate is set to 3e-5 and the batch size is 32. We set the number of epochs to 5. The maximal answer length and document length are set to 20 and 512, respectively. We set the length of document stride to 128. All experiments are conducted on 4 Tesla P40 GPUs.

E Human Performance

We evaluate human performance on both the in-domain test set and challenge test set. We randomly sample two hundred instances from the in-domain test set, and three hundred instances from the challenge test set. We ask crowdworkers to provide answers to the questions in the sampled instances. Then, we use EM and F1-scores of these annotated answers as human performance.

F Different Prediction Rate

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

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

where, $f(\theta;q)$ denotes the prediction of the trained MRC model $f(\theta)$. Q represents a set of pairs of

Types	# of changes (%)	# of same (%)
WR	1 (12.50)	7 (87.50)
RF	0 (00.00)	4 (100.00)
SS	6 (17.14)	29 (82.85)
AD	7 (23.33)	23 (76.66)
CO	7 (30.43)	16 (69.56)

Table 9: Distributions of paraphrases and DPRs.

Question Types	EM	F1	#
Company abbreviations	0	31.15	18
Pledgee	80.76	89.96	26
Pledgor	0	24.62	25
The pledge amount	18.36	53.84	98
Others (e.g. pledge date)	47.91	58.97	48
All	28.83	54.05	215

Table 10: The performance of RoBERTa$_{large}$ on the five topics in the domain of financial reports.

original question q and paraphrased question $q\prime$ in dataset D, and $\mathbb{1}[*]$ is an indicator function. A high DPR score means that the MRC model is overly sensitive to the paraphrased question $q\prime$, otherwise insensitive.

G Experimental Analysis

G.1 Over-sensitivity Analysis

We further analyze the prediction results to figure out what kind of paraphrases lead to different predictions. Five types of paraphrasing phenomena have been found, including (1) word reordering (WR), (2) replacement of function words (RF), (3) substitution by synonyms (SS), (4) inserting or removing content words (AD), and (5) more than one previously defined types happen in one paraphrase (CO). We randomly sample one hundred instances from the over-sensitivity subset and analyze the changes of the predictions by ERNIE (Sun et al., 2019). As shown in Table 9, most of changed predictions come from AD and CO. This analysis suggests that the models are sensitive to the changes of content words.

G.2 Generalization Analysis

In previous section, we have already analyzed the behaviors of baseline systems on education domain. In this section, we conduct analysis on financial domain. The data of financial domain contains management changes and equity pledge. The performance of RoBERTa$_{large}$ 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 about asking person names. By contrast,

the model performs worse on equity pledge. We classify the instances of equity pledge into five sets according to the question types. Table 10 shows the performance of RoBERTa$_{large}$ 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. By contrast, the model performs better on the questions about amount and date, since the model has already learnt relevant knowledge in the training set.