Are Machine Reading Comprehension Systems Robust to Context Paraphrasing?

Anonymous ACL submission

Abstract

Investigating the behaviour of pre-trained Machine Reading Comprehension (MRC) models under various types of test-time perturbations can shed light on the enhancement of their robustness and generalisation capability, despite the superhuman performance they have achieved on existing benchmark datasets. In this paper, we study the robustness of contemporary MRC systems to context paraphrasing, i.e., whether these models are still able to correctly answer the questions once the reading passages have been paraphrased. To this end, we systematically design a pipeline to semi-automatically generate perturbed MRC instances which ultimately lead to the creation of a paraphrased test set. We conduct experiments on this dataset with six state-of-the-art neural MRC models and we find that even the minimum performance drop of all these models exceeds 28%, whereas human performance remains high. These results demonstrate that the existing high-performing MRC systems are still far away from real language understanding.

Question: How much did Edison offer Tesla to redesign a motor and generators?
Original Context: [. . .] In 1885, he said that he could redesign Edison's inefficient motor and generators, making an improvement in both service and economy. According to Tesla, Edison remarked, "There's fifty thousand dollars in it for you—if you can do it." [. . .] **Prediction:** fifty thousand dollars
Paraphrased Context: [. . .] In 1885 he said that he could redesign Edison's inefficient motors and generators, making progress in both the ministry and the economy. According to Tesla, Edison commented, "There's fifty thousand dollars in it for you if you can do it. [. . .] **Prediction:** US$10 a week **Prediction by a human:** fifty thousand dollars

Figure 1: A perturbed test example in which Indonesian was used to back-translate (i.e., paraphrase) the context. The RoBERTa-large model originally provides the correct answer, but is confused by the paraphrased context.

1 Introduction

Machine reading comprehension (MRC), the task of automatically reading a passage of text and answering related questions, serves as an important testbed for evaluating various Natural Language Understanding (NLU) capabilities of computer systems (Chen, 2018). While neural MRC systems approach or even surpass human performance on benchmark datasets (Devlin et al., 2019; Lan et al., 2020; He et al., 2021), it remains uncertain whether they can indeed solve the MRC task (Jiang and Bansal, 2019; Schlegel et al., 2020; Sugawara et al., 2020; Wu et al., 2021b; Lai et al., 2021; Sugawara et al., 2022; Shinoda et al., 2023). In particular, recent studies have shown that instead of performing consistently well, contemporary models are brittle under various test-time perturbations (Zhou et al., 2020; Ribeiro et al., 2020; Si et al., 2021; Wu et al., 2021a; Lai et al., 2021; Schlegel et al., 2021; Yan et al., 2022). This raises the question of the suitability of existing gold standard datasets to establish a model's robustness and the need to improve the reliability of these MRC systems.

Paraphrase understanding plays a role in measuring the robustness and generalisation ability of MRC models. Intuitively, a trustworthy MRC system should demonstrate robust generalisation on paraphrased contexts and/or questions, i.e., those that convey the same semantic meaning using different surface forms. Previous researchers investigated an MRC model's robustness to question and partial-context level paraphrasing (Gan and Ng, 2019; Lai et al., 2021; Wu et al., 2021a). By paraphrasing questions or several sentences in the passages from SQuAD 1.1 (Rajpurkar et al., 2016) and

assessing model performance on the paraphrased test sets, they concluded that some MRC models might be vulnerable to paraphrasing-oriented attacks, which might be due to the large proportion of examples that can be characterised by trivial features (e.g., lexical overalap between the given question and context) that models learn to exploit during training (Lai et al., 2021).

The reading comprehension task assesses a model's real understanding of a given context, i.e., a *passage*. Though the findings in the work of Lai et al. (2021) and Wu et al. (2021a) provide insights into the weaknesses of MRC datasets to benchmark partial-context paraphrasing understanding, their designed perturbation methods are limited in their range, therefore the generated perturbed examples might not suitably estimate a model's robustness in the real-world test setting. Furthermore, it is not clear whether the modifications introduced as part of the perturbations changed the meaning of the original context. Therefore, to precisely reveal the capability of existing gold standard datasets to benchmark paraphrase understanding, we argue that it is crucial to examine the robustness of MRC systems to paraphrasing the whole context as well.

In this paper, our aim is to evaluate how well current reading comprehension systems generalise to a modified benchmark in which all contexts were paraphrased while preserving the same meaning and thus keeping the same gold standard answer. Different from prior robustness assessment research (Gan and Ng, 2019; Wu et al., 2021a; Yan et al., 2022), we design a pipeline to generate and identify perturbations of MRC examples that demonstrate the lack of robustness of MRC systems to context paraphrasing (see Figure 1 for an example). This proposed evaluation framework leads to the construction of a paraphrased test set drawn from the original SQuAD 1.1 benchmark (Rajpurkar et al., 2016). Results of our experiments show that the performance of six state-of-the-art MRC models on our created dataset is substantially lower, indicating that existing MRC datasets might insufficiently benchmark context paraphrasing understanding. This suggests that there is a need to create gold standard datasets in which context paraphrasing challenges are sufficiently represented.

2 Experiment Setup

MRC Dataset. In this paper, we investigated an extractive English MRC dataset SQuAD 1.1 (Rajpurkar et al., 2016) (License: CC-BY 4.0) due to its simplicity and the fact that it is the dataset on which current MRC models have already achieved superhuman performance, hence allowing us to focus on analysing the robustness of models to context paraphrasing. In the SQuAD 1.1 development set, there are 2,067 contexts and 10,570 questions.

Models. We chose the following models for the task of machine translation and reading comprehension, respectively.

Machine translation: We used the neural translation models provided by OPUS-MT (Tiedemann and Thottingal, 2020) which are based on the popular Marian-Neural Machine Translation framework (Junczys-Dowmunt et al., 2018) pre-trained on the OPUS (Tiedemann, 2012) multilingual corpus.

Reading comprehension: We selected the RoBERTa-large model (Liu et al., 2019) to generate the paraphrased test set mainly due to its impressive performance (93.1% F1) on the original development set of SQuAD 1.1 (Rajpurkar et al., 2016). In the final evaluation stage, we used multiple strong MRC models including BERT (Devlin et al., 2019), DistilBERT (Sanh et al., 2019), ALBERT (Lan et al., 2020), SpanBERT (Joshi et al., 2020) and DeBERTa (He et al., 2021), to comprehensively demonstrate the challenge posed by our created dataset. We fine-tuned these pre-trained language models on the training set of SQuAD 1.1 (Rajpurkar et al., 2016) and evaluated them on each of the original and perturbed test sets by making use of HuggingFace's *Transformers* library (Wolf et al., 2020). Model details and the hyperparameters used in model fine-tuning are shown in Appendix A.

3 Context Paraphrasing-Oriented Challenge Set Generation

In this section, we describe our methodology for generating semantics-preserving context-paraphrased datasets. Four steps are involved in the perturbation pipeline, which are detailed below.

3.1 Automatic Context Paraphrasing

We explored paraphrasing the reading passages in the development set of an MRC dataset using a back-translation approach, by which each sentence in the context is translated from a source language (English) to a pivot language and then back to the source language. We identified a total of twelve languages across five language families as the pivot language, informed by their number of speakers

Language	Performance (EM/F1)	
	Original	Paraphrased
Chinese	91.07/94.38	80.60/86.05$_{-8.83}$
Hindi	92.50/95.29	70.89/76.11$_{-20.13}$
Spanish	89.80/93.99	87.75/92.51$_{-1.57}$
French	90.26/94.24	87.20/92.02$_{-2.36}$
Russian	90.69/94.38	85.29/90.27$_{-4.35}$
German	89.95/94.22	87.51/92.29$_{-2.05}$
Italian	90.07/94.13	86.85/91.99$_{-2.27}$
Dutch	89.43/93.98	86.98/92.08$_{-2.02}$
Swedish	89.70/94.04	87.26/92.16$_{-2.0}$
Indonesian	91.07/94.53	84.39/89.54$_{-5.28}$
Vietnamese	91.73/95.06	77.14/82.89$_{-12.8}$
Finnish	91.20/94.62	85.15/90.22$_{-4.65}$

Table 1: The performance (%) of RoBERTa-large (Liu et al., 2019) on the original and paraphrased test sets generated using 12 pivot languages across five language families. Values in smaller font are changes in F1 (%) relative to the original performance of the model.

(Eberhard et al., 2022) and the performance of their associated pre-trained neural translation models (Tiedemann and Thottingal, 2020). After obtaining the paraphrases, we keep only those where all annotated answers can still be found in the paraphrased context. The original contexts of those paraphrases were discarded from the development set, to keep it aligned with the modified test set aligned and the performance comparable.

3.2 Preliminary Evaluation

As presented in Section 3.1, we generated perturbed test subsets (one for each of the 12 pivot languages) in which contexts were paraphrased using back-translation, and their corresponding original versions. Then, we examined the performance of a strong MRC model, RoBERTa-large (Liu et al., 2019), on these datasets, as demonstrated in Table 1. It can be seen from Table 1 that paraphrasing the contexts using different pivot languages caused various degrees of degradation in terms of the performance of the RoBERTa-large model. Nonetheless, we cannot simply conclude that this indicates the vulnerability of MRC models to the context paraphrasing attack as it is unclear whether these context paragraphs were indeed *paraphrased*, i.e., remain semantically equivalent while lexical/syntactic features were changed. Therefore, we manually verified the validity of the perturbed MRC instances in the next step.

3.3 Human Evaluation

With the aim of studying the lack of robustness of MRC models to context paraphrasing, from each generated perturbed test set, we identified MRC examples on which the RoBERTa-large model (Liu et al., 2019) predicts a wrong answer span whereas it provides the correct answer given the original passage. Afterwards, we randomly sampled 10% examples from each perturbed test set; this resulted in a total of 247 candidate examples, based on which human performance was assessed. A candidate perturbed MRC example has the ability to demonstrate the vulnerability of a model to context paraphrasing, if the model makes a wrong prediction on the paraphrased context paragraph, but a human can answer the question correctly. We refer to such candidates as *suitable* examples. Out of 247 examples, we identified 31 as suitable. The identification process is detailed in Appendix B.

3.4 Paraphrased Test Set Generation

While human evaluation enables us to identify suitable paraphrased MRC instances precisely, it requires significant human annotation effort. Therefore, we experimented with multiple classifiers to automatically determine whether a perturbed MRC example is suitable. We trained and evaluated the classifiers on our 247 annotated examples with input features that were calculated by TAACO (Crossley et al., 2019), a tool that measures various linguistic features of the passage such as lexical density and adjacent sentence overlap. The details of our model training and evaluation are reported in Appendix C. The best-performing model, Random Forest (Ho, 1995) (1.0 precision on predicting suitable example), was then applied on all candidate perturbed instances, 120 of which were classified as suitable. To increase this number, we further performed chained back-translation in which two languages were used to paraphrase the contexts. Through this step, 179 additional suitable examples were obtained (from 299 paraphrased contexts). For multiple paraphrased contexts that correspond to the same original passage, we only kept the one with the highest average question–context lexical overlap (Shinoda et al., 2021). Our final paraphrased test set contains 119 contexts and 127 questions in total.

Model	Original (EM/F1)	Paraphrased (EM/F1)
RoBERTa-large	100/100	0/10.49$_{-89.51}$
BERT-large	81.89/85.72	43.31/48.5$_{-43.42}$
DistilBERT-base	70.08/78.9	40.16/48.75$_{-38.21}$
SpanBERT-large	80.32/85.52	47.24/53.88$_{-37}$
ALBERT-xxlarge-v1	92.91/94.98	52.76/58.53$_{-38.38}$
DeBERTa-large	88.98/93.14	59.06/66.45$_{-28.66}$

Table 2: The performance (%) of the fine-tuned MRC models on the original and the paraphrased test set.

4 Results and Discussion

4.1 Evaluation

We assessed the performance of six state-of-the-art MRC models on the newly created challenge set, as shown in Table 2. The table shows that all the evaluated neural language models demonstrated poor generalisation to our generated test set. RoBERTa-large suffered the largest performance drop of 89.51%—this is within our expectation since its errors were used to identify suitable examples. For the other five model architectures, the relative changes were smaller than that of RoBERTa-large, but still very noticeable with over 28% performance decrease. This demonstrates the poor capability of these reading comprehension systems to properly deal with the paraphrased contexts. Apart from RoBERTa-large, the performance of all five MRC models remained consistent across both original SQuAD 1.1 development set (Rajpurkar et al., 2016) and paraphrased test set, with DeBERTa-large achieving the highest EM and F1 score, followed by ALBERT, SpanBERT, BERT and DistilBERT. Nonetheless, by comparing the relative drops between DistilBERT-base and some stronger models such as BERT-large, we found that the consistency in model performance rankings might not always apply to their robustness to context paraphrasing. While DistilBERT-base achieved the lowest original performance (78.9% F1), it presented higher robustness than the BERT-large model. Finally, the DeBERTa-large demonstrates the highest robustness to the designed perturbation in this work, with the smallest F1 decrease (28.66).

4.2 Error Analysis

In this section, we manually checked 50 perturbed examples on which the examined MRC models failed and identified two types of errors models made. Firstly, the paraphrasing of keywords in the sentence that is required to answer the question, along with some other lexical changes, often leads models to provide an incorrect answer. As can be seen in Figure 5 in Appendix D, though the models including RoBERTa, DistilBERT and SpanBERT can get the answer correctly over the original passage, they can be easily fooled once a word shared between the context and question is paraphrased, e.g., from “practiced" to “trained”. Moreover, another source of errors is the change in sentence structure. Figure 6 presents such an example, where it can be seen that almost all of the examined models (except BERT-large) failed in providing the correct answer “1804” due to the phrases in the paraphrased context being rearranged. This indicates that these high-performing systems might mostly rely on certain words matching between the question and the context to generate the answer, rather than truly understanding the passage. However, we also observed in a small proportion of examples that a mismatch between the answer provided by a model and the gold standard answer, does not necessarily mean that the model’s answer is erroneous: in some cases, the semantic meaning of the paraphrased context has changed or the model’s answer is arguably correct. This indicates that this work might be underestimating their robustness and there is scope for developing better techniques to remove the noise existing in the challenge set.

5 Conclusion

In this paper, we reveal the weaknesses of contemporary transformer-based reading comprehension systems to context paraphrasing. With the proposed perturbation framework, we generated a paraphrased challenge set, to which six high-performing MRC models generalise poorly. This informs us that to equip models with context paraphrasing understanding ability, there is a need to create large-scale benchmarks in which this reasoning challenge is precisely represented. Future work will include the design of better context paraphrasing approaches that not only generate large-scale and clean paraphrased datasets, but also can be applied to more complex MRC benchmarks to embed paraphrase understanding along with other NLU challenges into them. We also plan to investigate the behaviour of MRC models after re-training on those perturbed examples.

Limitations

The work presented in this paper has a few limitations. First, there is a need to examine other more challenging MRC datasets apart from SQuAD 1.1. This requires the development of better context paraphrasing techniques that can maintain the intended NLU challenges while paraphrasing the reading passages, which poses significant technical challenges. Further, since we performed sentence-level context paraphrasing and most of the questions in SQuAD 1.1 only require a single sentence to answer (Min et al., 2018), we regard a context as paraphrased as long as the meaning of the answer sentence has not been changed. However, there is scope for document-level paraphrasing methods that retain the meaning of the whole original context and also confuse the MRC models. In addition, we only focussed on the extractive MRC task and therefore it is necessary to expand the study to other types of MRC such as multiple choice and generation task. Finally, the classifier we developed to automatically identify examples that are appropriate for robustness assessment features a low recall. To improve it, we might consider establishing a more suitable example identification scheme, manually labelling more examples, and mining other linguistic features about the original and the paraphrased reading context as well.

References

Danqi Chen. 2018. *Neural Reading Comprehension and Beyond*. Ph.D. thesis, Stanford University.

Scott A. Crossley, Kristopher Kyle, and Mihai Dascalu. 2019. The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, 51(1):14–27.

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.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2022. Ethnologue: Languages of the World. Available at https://www.ethnologue.com.

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, Florence, Italy. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In *International Conference on Learning Representations*.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Yichen Jiang and Mohit Bansal. 2019. Avoiding reasoning shortcuts: Adversarial evaluation, training, and model development for multi-hop QA. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2726–2736, Florence, Italy. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Yuxuan Lai, Chen Zhang, Yansong Feng, Quzhe Huang, and Dongyan Zhao. 2021. Why machine reading comprehension models learn shortcuts? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 989–1002, Online. 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 *International Conference on Learning Representations*.

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. *CoRR*, abs/1907.11692.

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

OpenAI. 2023. Gpt-4 technical report.

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*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019*.

Viktor Schlegel, Goran Nenadic, and Riza Batista-Navarro. 2021. Semantics altering modifications for evaluating comprehension in machine reading. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13762–13770.

Viktor Schlegel, Marco Valentino, Andre Freitas, Goran Nenadic, and Riza Batista-Navarro. 2020. A framework for evaluation of machine reading comprehension gold standards. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5359–5369, Marseille, France. European Language Resources Association.

Kazutoshi Shinoda, Saku Sugawara, and Akiko Aizawa. 2021. Can question generation debias question answering models? a case study on question–context lexical overlap. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 63–72, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kazutoshi Shinoda, Saku Sugawara, and Akiko Aizawa. 2023. Which shortcut solution do question answering models prefer to learn? *Proceedings of the AAAI Conference on Artificial Intelligence*.

Chenglei Si, Ziqing Yang, Yiming Cui, Wentao Ma, Ting Liu, and Shijin Wang. 2021. Benchmarking robustness of machine reading comprehension models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 634–644, Online. Association for Computational Linguistics.

Saku Sugawara, Nikita Nangia, Alex Warstadt, and Samuel Bowman. 2022. What makes reading comprehension questions difficult? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6951–6971, Dublin, Ireland. Association for Computational Linguistics.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8918–8927. AAAI Press.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Winston Wu, Dustin Arendt, and Svitlana Volkova. 2021a. Evaluating neural model robustness for machine comprehension. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2470–2481, Online. Association for Computational Linguistics.

Yulong Wu, Viktor Schlegel, and Riza Batista-Navarro. 2021b. Is the understanding of explicit discourse relations required in machine reading comprehension? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3565–3579, Online. Association for Computational Linguistics.

Jun Yan, Yang Xiao, Sagnik Mukherjee, Bill Yuchen Lin, Robin Jia, and Xiang Ren. 2022. On the robustness of reading comprehension models to entity renaming. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 508–520, Seattle, United States. Association for Computational Linguistics.

Xiaorui Zhou, Senlin Luo, and Yunfang Wu. 2020. Co-attention hierarchical network: Generating coherent

long distractors for reading comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9725–9732.

A Hyperparameters of the Neural Reading Comprehension Models

Table 3 shows the hyperparameters used to fine-tune the pre-trained MRC models in this work.

Model/$_{Parameters(M)}$	d	b	lr	ep
RoBERTa-large$_{(355)}$	384	16	3e-5	2.0
BERT-large$_{(340)}$	384	12	3e-5	2.0
DistilBERT-base$_{(65)}$	384	8	3e-5	3.0
SpanBERT-large$_{(340)}$	512	4	2e-5	4.0
ALBERT-xxlarge-v1$_{(223)}$	384	12	3e-5	2.0
DeBERTa-large$_{(350)}$	384	4	3e-6	3.0

Table 3: The hyperparameters used to fine-tune each pre-trained MRC model (with number of parameters used in them). d is the size of the token sequence fed into the model, b is the training batch size, lr is the learning rate, and ep is the number of training epochs. We used stride = 128 for documents longer than d tokens.

B Human Annotation

This Appendix details the process of identifying the MRC examples that are suitable for context paraphrasing oriented robustness assessment through human annotation. We recruited two student annotators in total, one is the postgraduate researcher in Materials and another is a postgraduate student in Statistics. Before starting the annotation task, we asked both annotators to check a few examples and report the average time they spent for annotating each example. Based on this, We paid the annotators for their work by offering them coupons with a value of 20 pence for each example they annotate.

For the sampled 247 candidate instances, we first asked one annotator to answer questions based on the paraphrased contexts. The annotator was required to select the shortest span in the context that answered the question and was allowed to leave the answer as blank if the question is not answerable anymore. Furthermore, we asked the annotator to mark the questions that the annotator believes can be answerable based on the paraphrased context. Full text of instructions given to the annotator can be seen in Figure 2. After this, for the examples that the annotator is confident can be answerable, we compared the answer span provided by the annotator with the ground truth answers and labelled it as suitable or unsuitable based on the criteria described in Section 3.3. Since the annotation process is time-demanding and also for the purpose of evaluating the capabilities of the recently developed GPT-4 model (OpenAI, 2023), we used GPT-4 as the second annotator to answer questions and obtained the label for the 247 perturbed examples. We then measured the inter-annotator agreement by computing the Cohen's kappa coefficient, which is 0.32 and thus indicates a moderate agreement between the human annotator and the GPT-4 model. For the 81 examples that the two annotators share a disagreement, we presented them to the third annotator and provided them the label that agreed by the majority of annotators. We found that on around 80.25% (65 out of 81) examples, the first and the third human annotator shared the same view, which might indicates that GPT-4 still has a limitation on handling this annotation task compared with the human annotators. Importantly, we observed that while GPT-4 performs generally well on the answerable questions (see Figure 3 as an example), it demonstrates relatively unsatisfactory performance when handling those unanswerable examples. For instance, it can be seen from Figure 4 that since "sacks" (a technical term in American football) was wrongly paraphrased to "bags", the paraphrased context does not make it possible to answer the question anymore. While both human annotators regard this question as unanswerable, GPT-4 still provides the answer. This suggests that there is still a need to improve the reliability of the high-performing GPT-4 model.

C Identification of MRC Examples for Robustness Assessment

To avoid significant human annotation efforts, we attempted to automatically classify whether a reading comprehension example is qualified to demonstrate the lack of robustness of MRC models to context paraphrasing. We designed a classification pipeline which involves data standardisation, features selection and random oversampling, and experimented with ten classifiers. We used grid search to find the optimal hyperparameters. Our best-performing model, Random Forest (with 40 selected features), achieved 90% accuracy and 1.0 precision score on predicting suitable examples.

Thanks for contributing to this project! Your task is to read each given context and answer a question about it. We will compare the answer you provide with the ground truth answers to determine the human answerability of these questions, and then screen out the examples that are suitable for the robustness assessment of reading comprehension systems. When you are answering the questions: (1) If you meet a question that you truly think you can answer it based on the given context, then select the shortest continuous span in the context as it's answer and mark this type of question. (2) If you meet a question that could not be answered exactly but there were some key words in the question that allowed you to give the best answer available, still, select the shortest continuous span from the context as it's answer. (3) If you meet a question that is completely unanswerable, leave the answer as blank.

Figure 2: Instructions for the annotation task.

Paragraph: The Mongols attacked Samarkand using captured enemies as body shields. After several days only a few remaining soldiers, loyal supporters of the Shah, held out in the citadel. After the fortress fell, Genghis supposedly reneged on his surrender terms and executed every soldier that had taken arms against him at Samarkand. The people of Samarkand were ordered to evacuate and assemble in a plain outside the city, where they were killed and pyramids of severed heads raised as a symbol of victory. Ata-Malik Juvayni, a high official in the service of the Mongol empire, wrote that in Termez, on the Oxus, "all the people, both men and women, were driven out onto the plain, and divided in accordance with their usual custom, then they were all slain".
Paraphrased Paragraph: Mongols attacked Samarkand using captured enemies as body armor. After a few days, just a few remaining soldiers, the faithful adherents of the Shah, were in the fort. After the fort fell, Genghis was supposed to give up his surrender terms and execute all the soldiers who had taken arms against him in Samarkand. The inhabitants of Samarkand were ordered to evacuate and assemble on a plain outside the city where they were killed and the pyramids of the broken ends were raised to symbol of victory. Ata-Malik Juvayni, a high official of the Mongol Empire, wrote that in Termez Oxus, "all men and women were driven to the plain and divided according to the usual custom, and then all were killed".
Question: What did the Mongols shield themselves with while attacking Samarkand?
GPT-4's Answer: The Mongols attacked Samarkand using captured enemies as body armor, as mentioned in the context.

Figure 3: GPT-4's response on a paraphrased reading comprehension example.

D Perturbed Examples Demonstration

We present two perturbed examples from the constructed challenge set on which multiple strong MRC models demonstrated unsatisfactory generalisation, as shown in Figure 5 and Figure 6, respectively.

Paragraph: The Panthers defense gave up just 308 points, ranking sixth in the league, while also leading the NFL in interceptions with 24 and boasting four Pro Bowl selections. Pro Bowl defensive tackle Kawann Short led the team in sacks with 11, while also forcing three fumbles and recovering two. Fellow lineman Mario Addison added 6½ sacks. The Panthers line also featured veteran defensive end Jared Allen, a 5-time pro bowler who was the NFL's active career sack leader with 136, along with defensive end Kony Ealy, who had 5 sacks in just 9 starts. Behind them, two of the Panthers three starting linebackers were also selected to play in the Pro Bowl: Thomas Davis and Luke Kuechly. Davis compiled 5½ sacks, four forced fumbles, and four interceptions, while Kuechly led the team in tackles (118) forced two fumbles, and intercepted four passes of his own. Carolina's secondary featured Pro Bowl safety Kurt Coleman, who led the team with a career high seven interceptions, while also racking up 88 tackles and Pro Bowl cornerback Josh Norman, who developed into a shutdown corner during the season and had four interceptions, two of which were returned for touchdowns.
Paraphrased Paragraph: The Panthers' defense gave up only 308 points, ranking sixth in the league, while leading the NFL in the interceptions with 24 and boasting four Pro Bowl selections. Pro Bowl Kawann Short's defensive fight led the team in bags with 11, while also forcing three beats and recovering two. Mario Addison added 61⁄2 bags. The Panthers line also featured Jared Allen, a 5-time professional player who was the NFL's active career bag leader with 136, as well as Kony Ealy, who had 5 bags in just 9 starts. Behind them, two of the Panthers were also selected to play in the Pro Bowl: Thomas Davis and Luke Kuechly. Davis compiled 5 1⁄2 bags, 4 forced shots, and 4 interceptions, while Kuechly led the team in combats (118) forced two shots, and intercepted four passes of himself. Carolina's high school presented Pro Bowl Kurt Coleman's security, who led the team with a career of seven interceptions, while accumulating 88 attacks and Pro Bowl cornerback Josh Norman, who developed in a closed corner during the season and had four interceptions, two of which were returned for touchdowns.
Question: Who registered the most sacks on the team this season?
GPT-4's Answer: Kawann Short

Figure 4: Demonstration of the limitation of GPT-4 model in the task of MRC. Though the question becomes unanswerable, GPT-4 still generates the answer.

Paragraph: The Panthers used the San Jose State practice facility and stayed at the San Jose Marriott. The Broncos practiced at Stanford University and stayed at the Santa Clara Marriott.
Paraphrased Paragraph: Panthers use the San Jose State practice facility and remain in San Jose Marriott. Broncos trained at Stanford University and stayed at Santa Clara Marriott.
Question: Where did the Broncos practice for the Super Bowl?
Original Prediction: Stanford University **Prediction Under Context Paraphrasing:** Santa Clara Marriott

Figure 5: An example of RoBERTa-large, DistilBERT-base and SpanBERT-large fail to get the answer correct under the context paraphrasing perturbation.

Paragraph: Throughout the 18th century, Enlightenment ideas of the power of reason and free will became widespread among Congregationalist ministers, putting those ministers and their congregations in tension with more traditionalist, Calvinist parties.:1–4 When the Hollis Professor of Divinity David Tappan died in 1803 and the president of Harvard Joseph Willard died a year later, in 1804, a struggle broke out over their replacements. Henry Ware was elected to the chair in 1805, and the liberal Samuel Webber was appointed to the presidency of Harvard two years later, which signaled the changing of the tide from the dominance of traditional ideas at Harvard to the dominance of liberal, Arminian ideas (defined by traditionalists as Unitarian ideas).:4–5:24
Paraphrased Paragraph: During the 18th century, the congregation ministers spread ideas about the power of reason and free will and put these ministers and their communities in tension with traditional Calvinist parties.:1.4 When the Hollis professor of divinity David Tappan died in 1803, the President of Harvard Joseph Willard died a year later, a battle broke out in 1804 for their successors. Henry Ware was elected president in 1805, and the liberal Samuel Webber was appointed president of the Harvard presidency two years later, which changed the tide of dominance of traditional ideas at Harvard to the dominance of liberal, Arminian ideas (defined by traditionalists as unitary ideas).:4
Question: In what year did Harvard President Joseph Willard die?
Original Prediction: 1804 **Prediction Under Context Paraphrasing:** 1803

Figure 6: Illustration of the brittleness of MRC systems when dealing with a syntactic form changed context.