Benchmarking Robustness of Machine Reading Comprehension Models

Chenglei Si[1] **, Ziqing Yang**[2]**, Yiming Cui**[2,3]**, Wentao Ma**[2]**, Ting Liu**[3]**, Shijin Wang**[2,4]
[1] University of Maryland, College Park, MD, USA
[2] Joint Laboratory of HIT and iFLYTEK Research (HFL), Beijing, China
[3] Research Center for SCIR, Harbin Institute of Technology, Harbin, China
[4] State Key Laboratory of Cognitive Intelligence, Beijing, China
[1]sichenglei1125@gmail.com
[2]{zqyang5,ymcui,wtma,sjwang3}@iflytek.com
[3]{ymcui,tliu}@ir.hit.edu.cn

Abstract

Machine Reading Comprehension (MRC) is an important testbed for evaluating models' natural language understanding (NLU) ability. There has been rapid progress in this area, with new models achieving impressive performance on various MRC benchmarks. However, most of these benchmarks only evaluate models on in-domain test sets without considering their robustness under test-time perturbations. To fill this important gap, we construct AdvRACE (Adversarial RACE), a new model-agnostic benchmark for evaluating the robustness of MRC models under six different types of test-time perturbations, including our novel superimposed attack and distractor construction attack. We show that current state-of-the-art (SOTA) models are vulnerable to these simple black-box attacks. Our benchmark is constructed automatically based on the existing RACE benchmark, and thus the construction pipeline can be easily adopted by other tasks and datasets. We will release the data and source codes to facilitate future work. We hope that our work will encourage more research on improving the robustness of MRC and other NLU models.[1]

1 Introduction

Machine reading comprehension refers to the task where the system is given a passage and corresponding questions, and it needs to predict the correct answer to the question based on the passage. MRC is considered as a challenging task because it requires the model to understand the text and even perform some types of reasoning in order to correctly answer the questions. To this end, many MRC benchmarks have been constructed with different domains, styles and languages (Rajpurkar et al., 2016; Lai et al., 2017; Campos et al., 2016; Dua et al., 2019b; Dasigi et al., 2019; Cui et al., 2019, *inter alia*). While most of these benchmarks use leaderboards to compare different models' performance on in-domain test sets, they have ignored the important aspect of evaluating models' robustness.

The research on robustness of MRC models can be generally categorised into two directions: generalization to out-of-domain distributions and robustness under test-time perturbations.

On the generalization to out-of-domain distributions, Talmor and Berant (2019) has investigated how well do MRC models trained on source MRC datasets generalize to unseen datasets. The MRQA workshop (Fisch et al., 2019) also hosted a shared task where MRC systems trained on the given training sets are evaluated on hidden test sets with different distributions.

On evaluating MRC models under test time perturbations, existing work (Jia and Liang, 2017; Gan and Ng, 2019) typically proposes a new adversarial attack method and evaluates models under this specific attack. However, robust models that can be reliably deployed in real-life applications should be able to perform well under various types of perturbations instead of just one. Hence in this work, we aim to construct a benchmark that consists of diverse types of test-time perturbations to allow for a more comprehensive evaluation of models' robustness.

Instead of constructing a new benchmark from scratch, we leverage on an existing MRC benchmark RACE (Lai et al., 2017), where we apply our proposed perturbations on the RACE test set to form a new set of adversarial test sets which we name as AdvRACE. We choose to construct our benchmark based RACE for the following reasons: 1) The format of multiple-choice MRC allows more types of attacks. For example, we generated new distractors to replace

[1]Work in progress. Contribution during the first author's internship at HFL.

the original ones as a novel way of attack. 2) RACE covers a diverse set of linguistic phenomena and reasoning types, and has been used widely for evaluation of NLU models. For example, recent pretrained language models like XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019) and ALBERT (Lan et al., 2020) have all used RACE to test their language representation ability. By constructing an adversarial version of RACE, we allow the community to compare and analyse their robustness easily.

We describe the desired properties of our AdvRACE benchmark and the rationales behind them: 1) We use an automated pipeline to construct all the test-time perturbations. This allows us to generate the benchmark efficiently and at minimal cost. It also allows use to apply this pipeline on other datasets to construct the corresponding adversarial test sets with little adaptation. 2) All of our perturbations are model-agnostic, meaning that they are not targeting any specific model. Thus it provides a fair comparison for all the models, including future new models. 3) Our attacks are black-box and do not require access to model parameters. This allows us to fix the benchmark without the need to generate new perturbations based on the models being evaluated. 4) Our perturbations are label-preserving. We will include human evaluation to ensure that the perturbed test set retains the original labels and are answerable for humans.[2]

Given these motivations, we adapt three existing types of adversarial attacks: distracting sentence insertion, character misspelling, and sentence paraphrasing. We also propose two new types of attacks that are designed for multiple-choice MRC: distractor extraction and distractor generation. Moreover, we also explore a new setting of adversarial attack: superimposed attacks, where multiple attacks are applied to the data at the same time. The details are described in Section 3.

In summary, our contributions are:

- We propose an automated pipeline to construct AdvRACE, a benchmark that evaluates MRC models under various test-time perturbations, which allows for a more comprehensive evaluation and analysis of models' robustness.
- We propose new attack methods on multiple-choice MRC by constructing new distractors through extraction and generation.
- We explore the setting of superimposed attacks where multiple attacks are applied simultaneously on the dataset.

[2]Human evaluation results will be updated later.

2 Related Work

Robustness of NLP models. Previous works have explored different methods to construct adversarial test sets to evaluate robustness of models. One type is white-box attack where perturbations are constructed based on models' gradients or parameters to exploit their weakness (Ebrahimi et al., 2018; Wallace et al., 2019). Another type is black-box attack where the perturbation is independent of the model being tested and is often based on heuristic rules or produced by designed models (Jia and Liang, 2017; Ribeiro et al., 2018; Iyyer et al., 2018). There are also works on human-in-the-loop adversarial example generation where instead of automatically applying perturbations on the original data, human annotators are employed to write new test data that can fool the models into making wrong predictions (Nie et al., 2019; Wallace et al., 2018).

New evaluation protocol. There are several previous attempts on better evaluating NLP models. To prevent models from exploiting spurious patterns instead of the linguistic capacities intended in the datasets, Kaushik et al. (2020) constructed counterfactually-augmented data, Gardner et al. (2020) constructed contrast sets, where in both cases annotators rewrite the input with minimal changes so that it accords with a counterfactual target label. They found that NLP models trained on original training sets perform poorly on these human-created test sets. These approaches are fundamentally different from our work in the sense that: 1) These approaches change the original labels, while we create label-preserving perturbations only. 2) These works employ human annotators to create the perturbations, while we automatically generate all the perturbations, which is much faster and cheaper.

Datasets focusing on robustness. Dua et al. (2019a) provided an evaluation server consisting of seven diverse MRC datasets so that models can be tested on a variety of reading phenomena. They also performed synthetic augmentation for the datasets to test models' robustness during test time.

This is spiritually similar to our work. However, their synthetic transformation techniques have several flaws: 1) Most of their techniques are only applicable to specific types of questions (e.g. binary choice questions). In fact, four of their five proposed techniques yield only small number of valid augmentations. 2) Their synthetic augmentations are not adversarial enough to exploit the model's weakness. This means that most of their methods fail to create a large enough difference to make the model fail on the perturbed data. This may be partially because that they only performed perturbations on the questions. In contrast, our proposed methods are applied to different components (e.g., passages, questions, distractors) of the dataset and at different levels (e.g. sentence and character levels), and all of them are effective in attacking the models. In a contemporary work, Tang et al. (2020) created a Chinese MRC dataset to test the robustness and generalization ability of models. Their robustness test set has similar motivation as ours, while their data are retrieved from databases or written by humans instead of automatically generated, and they only focused on rewriting the questions.

3 Methods

In this section, we describe in detail the six types of perturbations that we apply to the original RACE test set. The first three types are adapted from existing work, where we apply them on the multiple-choice MRC setting with some modifications. The last three types of adversarial perturbations are newly proposed in this work for attacking MRC models. A summary of all these perturbations is shown in Table 1. We also present examples of each perturbation in Table 3.

3.1 AddSent

Inspired by the original AddSent method proposed by Jia and Liang (2017), we adapt it to the multiple-choice setting. The motivation is to add distracting information that is similar to the question so that it can mislead models that rely largely on text matching. We use the following procedure to construct the perturbation:

1. We change all the nouns, named entities and numbers in the questions to their nearest word in GloVe (Pennington et al., 2014) embedding space with the same part of speech.
2. We replace adjectives, adverbs, verbs in the questions with their antonyms in WordNet (Fellbaum, 2000).
3. If no words are changed during Step 1 and 2, we randomly sample a question from the test set with the same question word (i.e., wh-words, or fill-in-the-blank) to replace the original question.
4. We randomly sample a distractor from the original three distractors of the question, and concatenate it with the altered question.
5. We insert the concatenated sequence into a random position of the passage.
6. We repeat Step 1 to 5 one more time with different replacement words from GloVe and WordNet, and using a different distractor.

3.2 CharSwap

It is shown in Belinkov and Bisk (2018) that NMT performance drops significantly when there are spelling errors in the data. We follow their CharSwap approach to swap two adjacent letters in a word without altering the first or last letters. Although they show that with more tokens altered, the performance gets worse, it is risky to apply such perturbation to all words in the dataset because it may impact the readability of the text and cause difficulty even for humans to perform well. As a result, we only apply the CharSwap perturbation to the following words: 1) The non-stopwords in the question. 2) Non-stopwords in the passage that have also appeared in the question and its corresponding options. This is based on the motivation that perturbing such words would prevent models from relying on keywords matching. There are a total of 7.1% words being altered using this method.

3.3 Paraphrase

Paraphrasing has been used to generate adversarial examples. Gan and Ng (2019) generated paraphrases of the questions in SQuAD as a way to attack MRC models. However, due to the different nature of SQuAD and RACE, a large proportion of RACE questions involve multi-sentence reasoning and it is difficult to obtain valid paraphrase suggestions from the passages in order to generate effective question paraphrases that preserve the original labels. Therefore, we adopt the syntactically controlled paraphrase network (SCPN) (Iyyer et al.,

Perturbation	**Perturbation Level**	**Applied Component**	**MCRC-specific**
AddSent	Sentence	Passage	No
CharSwap	Character	Passage + Question	No
Paraphrase	Sentence	Passage	No
Superimposed	Sentence + Character	Passage	No
Distractor Extraction	Sentence	Distractors	Yes
Distracor Generation	Sentence	Distractors	Yes

Table 1: Summary of our perturbations. MCRC-specific means whether the method is specific to the format of multiple-choice reading comprehension.

Test Set	**BERT**	**RoBERTa**	**XLNet**	**ALBERT**
Original	69.5	83.7	79.9	86.0
AddSent	30.0 *(-56.8%)*	57.3 *(-31.5%)*	51.4 *(-35.7%)*	57.8 *(-32.8%)*
CharSwap	48.8 *(-29.8%)*	69.4 *(-17.1%)*	63.4 *(-20.7%)*	73.0 *(-15.1%)*
Paraphrase	59.4 *(-14.5%)*	72.3 *(-13.6%)*	68.2 *(-14.6%)*	73.7 *(-14.3%)*
Superimposed	18.6 *(-73.2%)*	38.1 *(-54.5%)*	36.4 *(-54.4%)*	36.1 *(-58.0%)*
Distractor Extraction	32.0 *(-54.0%)*	47.5 *(-43.2%)*	42.9 *(-46.3%)*	50.7 *(-41.0%)*
Distractor Generation	55.5 *(-20.1%)*	67.7 *(-19.1%)*	63.8 *(-20.2%)*	69.9 *(-18.7%)*
Average	40.7 *(-41.4%)*	58.7 *(-29.9%)*	54.4 *(-32.0%)*	60.2 *(-30.0%)*

Table 2: Attack results on different models. *Numbers* in brackets are the percentage drop in performance.

2018). This model allows us to generate paraphrases of a sentence based on a given syntactic parsing template, and it has been shown to be more effective in attacking compared to other paraphrase models such as back-translation. Through preliminary experiments, we find that: 1) Generating paraphrases of questions is difficult as any syntactic transformation of the questions may cause a misfit between question and options, especially for fill-in-the-blank types of questions. Moreover, paraphrases sometimes change or miss keywords from the questions which may not preserve the original correct answer. 2) Generating paraphrases with a different syntactic parse of the original sentence sometimes results in unnatural and incoherent sentences, due to the differences in syntactic parse. Therefore, we propose to adapt SCPN on multiple-choice MRC in the following way: 1) We generate paraphrases of sentences in the passage, using its original syntactic parse template. 2) We only paraphrase sentences in the passage that have lexical overlap (non-stopwords) with the question or options, and we keep the questions unchanged. This results in 47% of sentences in the passages being paraphrased.

3.4 Superimposed

Apart from applying one single attack on the original dataset, we also explore a new setting - superimposed attack, where multiple types of perturbations are applied to the same data simultaneously. For this superimposed attack, we first apply paraphrase on the passages, then apply CharSwap (this time without modifying the questions) on the paraphrased passages. Lastly, we perform AddSent by adding distracting information to the perturbed passages. The motivation of this attack is to test how models handle complex situations where defense against one simple attack is not sufficient.

3.5 Distractor Extraction

The previous three methods are modified from existing attack methods. Due to the special format of multiple-choice MRC, distractors naturally play an important part in increasing the difficulty of the dataset. Hence, a new type of attack towards multiple-choice MRC is to construct new sets of distractors to replace the original ones.

We present two different methods of constructing new distractors for multiple-choice MRC: extraction and generation. For distractor extraction, we aim to extract spans from the passages as new distractors. This is based on the motivation that:

Perturbation	Passage (shortened)	Question & Answer	Distractors
Original	Homeschooling is a legal choice for parents in developed countries to provide their children with a learning environment at home. Homeschooling can also be a choice for families living in remote locations, living abroad, and to allow for more traveling. Also many young athletes and actors are taught at home, where a coach or tutor is with the child for many years and then knows the child very well. In some places, an approved curriculum is required if children are to be home-schooled. In some cases a liberal arts education is provided.	**Q:** Which of the following is TRUE about homeschooling according to the text? **A:** Some parents have to homeschool their children when living abroad.	**1.** Homeschooling is still illegal in developed countries. **2.** Athletes and actors can not be home-schooled. **3.** There is no curriculum for homeschooled children.
AddSent	... In some places, an approved curriculum is required if children are to be home-schooled. ***Which of the leading is untruthful about homeschooling according to the document? Athletes and actors can not be home-schooled.*** In some cases a liberal arts education is provided. ***Which of the following is false about home-schooling according to the translation? Homeschooling is still illegal in developed countries.***	Same as **Original**	Same as **Original**
CharSwap	... ***Homescoholing*** can also be a choice for families ***livnig*** in remote locations, ***liivng aborad***, and to allow for more traveling . Also many young ***atlhetes*** and ***acotrs*** are taught at home, where a coach or tutor is with the child for many years and then knows the child very well . In some places, an approved ***cruriculum*** is required if ***cihldren*** are to be home-schooled. ...	**Q:** Which of the ***fololwing*** is TURE about ***hmoeschooling accroding*** to the ***txet***? **A:** Same as **Original**	Same as **Original**
Paraphrase	***Now home is a legal choice for parents in developed countries with a learning environment at home would be provided.*** Homeschooling can also be a choice for families living in remote locations, living abroad, and to allow for more traveling. ...	Same as **Original**	Same as **Original**
Superimposed	***Now home is a legal choice for parents in developed countries with a learning environment at home would be provided. Homecshooling*** can also be a choice for families ***livnig*** in remote locations, ***liivng arboad***, and to allow for more traveling. Also many young ***atheltes*** and ***acotrs*** are taught at home, where a coach or tutor is with the child for many years and then knows the child very well. In some places, an approved ***curriuclum*** is required if ***chilrden*** are to be home-schooled. ***Which of the leading is untruthful about homeschooling according to the document? Athletes and actors can not be home-schooled.*** In some cases a liberal arts education is provided. ***Which of the following is false about homeschooling according to the translation? Homeschooling is still illegal in developed countries.***	Same as **Original**	Same as **Original**
Distractor Extraction	Same as **Original**	Same as **Original**	**1.** allow for more traveling. **2.** an approved curriculum is required if children are to be home-schooled. **3.** most childhood education occurred within the family or community.
Distractor Generation	Same as **Original**	Same as **Original**	**1.** Homeschooling is a legal choice for parents. **2.** Homeschooling is an approved curriculum. **3.** There are many reasons for homeschooling.

Table 3: Examples of each perturbation applied on the same original test example. ***Italic*** parts are altered by our perturbations.

1) Many passages in RACE contain distracting information that is relevant to the question but different from the correct answer. 2) Such spans may be especially effective when attacking models that rely on text matching because they are directly extracted from the passage.

However, the key challenge is how do we extract such distracting spans from the passages. In order to solve this challenge, we propose a novel distractor extraction model. During training, we insert the correct answer of RACE into the passages and turn it into span-extraction format. The model is trained to select the answer span from the passage. During inference, we use the trained model to extract spans from the passages (without inserting the correct answer) for the given questions. In this way, the extracted spans with high probability can be considered as likely answers for the question. For post-processing, we select 3 distractors among the top 20 candidate spans that have low lexical overlap with each other and also low lexical overlap with the correct answer, so that they are more diverse and label-preserving. We used ALBERT (Lan et al., 2020) as the backbone model for span-extraction.

Note that some of the extracted spans are not well-formed, which means that they may not be syntactically coherent when put together with the question. However, this does not diminish their effectiveness on attacking the models.

3.6 Distractor Generation

Another way to construct new distractors is to directly generate them based on the passage and questions. This can be formed as a sequence-to-sequence problem where the input is the concatenation of passage and question, and the output is the distractor. There have been previous studies on distractor generation such as Gao et al. (2019) and Zhou et al. (2020) which adopt this sequence-to-sequence approach. However, none of these works has explored using the generated distractors as a way for attacking MRC models. We will use the cleaned data provided by Gao et al. (2019) where distractors with low semantic relevance with the passage are pruned. For details on the data cleaning process and the cleaned dataset's statistics, we refer readers to Gao et al. (2019). Unfortunately, we could not obtain the source codes from Zhou et al. (2020) which claimed to be SOTA for distratcor genration. Instead, we adopt UniLM (Dong et al., 2019), a pretrained model unifying NLU and NLG, which has achieved SOTA performance on several NLG tasks such as question generation.

We finetune UniLM on the training data for 10 epochs, and follow the decoding strategy in Gao et al. (2019). Specifically, use beam search to find the top k (beam size, we used $k = 50$) candidate distractors and select the top 3 among them such that the Jaccard distance between each pair is larger than 0.5. This is to ensure diversity of the 3 distractors.

4 Evaluation

We apply each of the perturbation in Section 3 on the original RACE test set, which results in six sets of new test sets, and they form our AdvRACE benchmark. Each of test set has 4,934 questions. For evaluation, models should be trained on the original RACE training set, tune hyperparameters on the original RACE dev set, and eventually test on our AdvRACE benchmark. We can use the average performance of the model across the six test sets in AdvRACE as a metric to compare different models' robustness, similar to how GLUE (Wang et al., 2019) is used for NLU evaluation.

To demonstrate how to use AdvRACE for evaluation and analysis, we evaluate four competitive models on the RACE leaderboard: BERT (Devlin et al., 2019), XLNet, RoBERTa, ALBERT. They are trained and tuned on the original RACE train and dev set, and we evaluate the model on both the original test set and our AdvRACE benchmark. The results are presented in Table 2.

By comparing the average percentage drop in performance (last row), we may conclude that BERT is the most vulnerable among these four models. Specifically, by comparing the performance under AddSent, we find that BERT suffered much more percentage drop in performance under AddSent compared to the other models.

Additionally, we also find that: 1) All the models degrade significantly on the perturbed test sets. 2) The superimposed attack is much more effective than the respective individual perturbation, but the resultant percentage drop is relatively smaller than the linear sum of the percentage drop of the three individual perturbations. 3) Distractor extraction is more effective than distractor generation in terms of attacking the models. This is probably due to the fact that models rely more on text matching (see Si et al., 2019).

5 Conclusion

In this work, we have constructed AdvRACE, a benchmark that evaluates MRC models under diverse types of test-time perturbations. In this benchmark, we also explored novel superimposed attack, distractor extraction and generation attacks, which are shown to be effective. We hope that AdvRACE can encourage future work on more robust NLU models and more effective defense methods.

References

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *ICLR*.

Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *ArXiv*, abs/1611.09268.

Yiming Cui, Ting Liu, Li Xiao, Zhipeng Chen, Wentao Ma, Wanxiang Che, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for chinese machine reading comprehension. In *EMNLP-IJCNLP*.

Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *EMNLP-IJCNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *NeurIPS*.

Dheeru Dua, Ananth Gottumukkala, Alon Talmor, Sameer Singh, and Matt Gardner. 2019a. Orb: An open reading benchmark for comprehensive evaluation of machine reading comprehension. *ArXiv*, abs/1912.12598.

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

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for nlp. In *ACL*.

Christiane Fellbaum. 2000. Wordnet : an electronic lexical database. *Language*, 76:706.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. *ArXiv*, abs/1910.09753.

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

Yifan Gao, Lidong Bing, Piji Li, Irwin King, and Michael R. Lyu. 2019. Generating distractors for reading comprehension questions from real examinations. In *AAAI*.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Quan Zhang, and Ben Zhou. 2020. Evaluating nlp models via contrast sets. *ArXiv*, abs/2004.02709.

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. In *EMNLP*.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *ICLR*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *EMNLP*.

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. In *ICLR*.

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. *ArXiv*, abs/1907.11692.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *ArXiv*, abs/1910.14599.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

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

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *ACL*.

Chenglei Si, Shuohang Wang, Min-Yen Kan, and Jing Jiang. 2019. What does bert learn from multiple-choice reading comprehension datasets? *ArXiv*, abs/1910.12391.

Alon Talmor and Jonathan Berant. 2019. Multiqa: An empirical investigation of generalization and transfer in reading comprehension. In *ACL*.

Hongxuan Tang, Jing Liu, Hongyu Li, Yu Hong, Hua Wu, and Hai-Feng Wang. 2020. Dureaderrobust: A chinese dataset towards evaluating the robustness of machine reading comprehension models.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Empirical Methods in Natural Language Processing*.

Eric Wallace, Pedro Rodriguez, S. L. Feng, Ikuya Yamada, and Jordan L. Boyd-Graber. 2018. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.

Xiaorui Zhou, Senlin Luo, and Yunfang Wu. 2020. Co-attention hierarchical network: Generating coherent long distractors for reading comprehension. In *AAAI*.