Towards Robust Neural Machine Reading Comprehension via Question Paraphrases

Ying Li
National Engineering Laboratory for
Brain-inspired Intelligence Technology and Application (NEL-BITA)
University of Science and Technology of China,
Hefei, China
nicole@baidu.com

Hongyu Li, Jing Liu
Baidu Inc.
Beijing, China
lihongyu04@baidu.com, liujing46@baidu.com

***Abstract*—In this paper, we focus on addressing the *oversensitivity* issue of neural machine reading comprehension (MRC) models. By *oversensitivity*, we mean that the neural MRC models give different answers to question paraphrases that are semantically equivalent. To address this issue, we first create a large-scale Chinese MRC dataset with high-quality question paraphrases generated by a toolkit used in Baidu Search. Then, we quantitively analyze the *oversensitivity* issue of the neural MRC models on the dataset. Intuitively, if two questions are paraphrases of each other, a robust model should give the same predictions. Based on this intuition, we propose a regularized BERT-based model to encourage the model give the same predictions to similar inputs by leveraging high-quality question paraphrases. The experimental results show that our approaches can significantly improve the robustness of a strong BERT-based MRC model and achieve improvements over the BERT-based model in terms of held-out accuracy. Specifically, the different prediction ratio (DPR) for question paraphrases of the proposed model decreases more than** 10%.

***Keywords*-machine reading comprehension; oversensitivity; question paraphrases;**

I. Introduction

Machine reading comprehension (MRC) requires machines to understand text and answer questions about the text, and it is an important task in natural language processing. With the increasing availability of large-scale labeled datasets for MRC ([1], [2], [3]) and the development of deep learning techniques ([8], [10], [11], [12], [13], [14], [15]), MRC has achieved remarkable advancements in the last few years.

Although a number of neural models obtains high held-out accuracy on several datasets, previous studies show that most of the complex neural models are not robust: different ways of phrasing the same question can often cause different answers.

Specifically, given a passage and two questions that are paraphrases of each other, a neural MRC model with high held-out accuracy may give different answers. As shown in Table I, *Question 1* and *Question 2* are paraphrases of each other, and we expect that a neural MRC model gives the same answer to these two questions. However, a BERT-based model that is one of the state-of-the-art MRC models, predicts two different answers. Additionally, the difference between *Question 3* and *Question 4* is just one question mark. Surprisingly, the BERT-based model again gives different predictions to these two questions.

These above examples suggest that the neural MRC models are very sensitive to similar inputs that are semantically equivalent. The *oversensitivity* of the neural MRC models may limit their applications to question answering systems or search engines, which require consistent predictions on various inputs. For example, the search engine users may use different ways to express the same information need. If the system provides different answers to the questions that are paraphrases of each other, it may hurt the user experiences.

Given the great variety of languages for semantically equivalent expressions, it is not surprise that previous work has investigated the use of paraphrases to machine reading comprehension or question answering systems. The previous work can be classified into three categories. The first one uses paraphrases in the context of neural question answering models ([18], [19]), and encourages the models to learn similar representations for the questions that are paraphrases of each other. Another category of work directly generates question paraphrases and applies the question paraphrases to a question answering module by scoring them, because the generated paraphrases often contain low-quality candidates ([17], [21], [22], [23], [24]). The third category of research mines high quality semantically equivalent adversarial rules to generate question paraphrases by involving human-in-the-loop [16].

Although the previous work tried to incorporate question paraphrases to improve the performance of the question answering systems, they did not explicitly address the *oversensitivity* issue. It is not clear to what extent the issue was addressed.

In this paper, we focus on addressing the *oversensitivity* issue of neural MRC models. We first create a large-scale Chinese MRC dataset with high-quality question paraphrases generated by a toolkit used in Baidu Search. Then, we quantitively analyze the *oversensitivity* issue of a BERT-based MRC model, that is one of the state-of-the-art neural MRC models. Intuitively, if two questions are paraphrases of each other, a robust model should give the same predictions. Based on this intuition, we propose a regularized BERT-based model by incorporating the high-quality question paraphrases. The experimental

978-1-7281-5014-7/19/$31.00 ©2019 IEEE

Table I
THE EXAMPLES OF THE OVERSENSITIVITY OF A BERT-BASED MRC MODEL.

Passage: 12月24号是平安夜，12月25号是圣诞节；分别相当于中国的大年29和大年30。(December 24th is the Christmas Eve, and December 25th is the Christmas; they are equivalent to the Spring Festival Eve and the Spring Festival in China.)
Question 1: 12月24日是什么日子 (What special day is December 24) Predicted Answer: 平安夜 (Christmas Eve)
Question 2: 12月24日是什么节日 (What holiday is December 24) Predicted Answer: 圣诞节 (Christmas)
Passage: 求一个数的立方根的运算方法，叫做开立方。它是立方的逆运算，最早在我国的九章算术中有对开立方的记载。由于任何实数均有唯一的立方与之对应且不存在两个实数的立方相等，故任何实数都存在且仅存在唯一的立方根。 (The method of finding the cube root of a number is called “kai li fang” in Chinese. It is the inverse of the cube operation. It was first recorded in the “Jiuzhang arithmetic” in our country. Since any real number and its cube is a unique pair and no cubes of two real numbers are equal, so the cube root of any real number exists and is unique.)
Question 3: 任何实数存在多少立方根 (How many cube roots does a real number have) Predicted Answer: 唯一 (Unique)
Question 4: 任何实数存在多少立方根? (How many cube roots does a real number have?) Predicted Answer: 开立方 (Cube root)

results show that our approaches can significantly improve the robustness of a strong BERT-based MRC model and achieve improvements over the BERT-based model in terms of held-out accuracy.

The contributions of this paper are three-folds:

- We create a large-scale Chinese MRC dataset with high-quality question paraphrases (see Section II-C). The dataset contains $85K$ passages and $242K$ questions, and each of the question has $5 \sim 10$ high-quality paraphrases.
- We quantitively analyze the *oversensitivity* issue of a BERT-based MRC model (see Section III-A). To the best of our knowledge, this is the first study of its kind, indicating a potential direction for future research.
- To address the *oversensitivity* issue, we propose a regularized BERT-based model by incorporating high-quality question paraphrases (see Section III-B). The experimental results show that our approaches can significantly improve the robustness of a strong BERT-based MRC model and achieve improvements over the BERT-based model in terms of held-out accuracy. Specifically, the different prediction ratio (DPR) for question paraphrases of the proposed model decreases more than 10% (see Section IV).

The remainder of this paper is organized as follows. Section II describes a strong BERT-based MRC model, and introduces a large-scale MRC dataset with high-quality question paraphrases generated by a toolkit used in Baidu Search. Section III quantitively analyzes the *oversensitivity* issue of a BERT-based MRC model, and proposes a regularized BERT-based model by incorporating the high-quality question paraphrases. In Section IV, we give the experimental results. Section V represents the related work. Section VI ends with conclusions.

II. BACKGROUND

A. A BERT-based MRC Model

Recently, the pre-trained language models ([12], [13], [14], [15]) have caused a stir in the MRC community. Among the pre-trained models, BERT [13], which uses Transformer encoder and trains a bidirectional language model, is one of the most successful models by far, presenting new state-of-the-art results in MRC. In this paper, we choose a BERT-based model as a baseline. Basically, a BERT-based MRC model has an encoding layer and an output layer.

BERT Encoding Layer: This layer uses Transformer encoder to model passages and questions. It takes as input passage P and question Q, and computes for each token a context-aware representation.

Specifically, given passage $P = \{p_i\}_{i=1}^{m}$ and question $Q = \{q_j\}_{j=1}^{n}$, we first pack them into a single sequence of length $m + n + 3$, i.e.,

$$S = [\langle \texttt{CLS} \rangle, Q, \langle \texttt{SEP} \rangle, P, \langle \texttt{SEP} \rangle],$$

where $\langle \texttt{SEP} \rangle$ is the token separating Q and P, and $\langle \texttt{CLS} \rangle$ the token for classification (will not be used in this paper). For each token s_i in S, we construct its input representation as

$$\mathbf{h}_i^0 = \mathbf{s}_i^{\text{tok}} + \mathbf{s}_i^{\text{pos}} + \mathbf{s}_i^{\text{seg}},$$

where $\mathbf{s}_i^{\text{tok}}$, $\mathbf{s}_i^{\text{pos}}$, and $\mathbf{s}_i^{\text{seg}}$ are the token, position, and segment embeddings for s_i, respectively. Tokens in Q share a same segment embedding $\mathbf{q}^{\text{seg}}$, and tokens in P share a same segment embedding $\mathbf{p}^{\text{seg}}$. Such input representations are then fed into L successive Transformer encoder blocks, i.e.,

$$\mathbf{h}_i^{\ell} = \text{Transformer}(\mathbf{h}_i^{\ell-1}), \;\; \ell = 1, 2, \cdots, L,$$

so as to generate deep, context-aware token representations for passages and questions. We refer readers to [13] for details. The final hidden states $\{\mathbf{h}_i^L\}_{i=1}^{m+n+3} \in \mathbf{R}^{d_1}$ are taken as the output of this layer.

Output Layer: We follow BERT and simply use a linear output layer, followed by a standard softmax operation, to predict answer boundaries. The probability of each token s_i to be the start or end position of the answer span is calculated as:

$$p_i^1 = \frac{\exp(\mathbf{w}_1^\top \mathbf{h}_i)}{\sum_j \exp(\mathbf{w}_1^\top \mathbf{h}_j)}, \ \ p_i^2 = \frac{\exp(\mathbf{w}_2^\top \mathbf{h}_i)}{\sum_j \exp(\mathbf{w}_2^\top \mathbf{h}_j)},$$

where $\mathbf{h}_i$ is the output of token s_i by the encoding layer, and $\mathbf{w}_1, \mathbf{w}_2 \in \mathbf{R}^{6d_1+6d_2}$ are trainable parameters. The training objective is the negative log-likelihood of the true start and end positions:

$$\mathcal{L}_1 = -\frac{1}{N}\sum_{j=1}^{N}(\log p^1_{y_j^1} + \log p^2_{y_j^2}),$$

where N is the number of examples in the dataset, and y_j^1, y_j^2 are the true start and end positions of the j-th example, respectively. At inference time, the span (a, b) where $a \leq b$ with maximum $p_a^1 p_b^2$ is chosen as the predicted answer.

B. A Question Paraphrase Generation Toolkit

There are various resources (e.g. PPDB [25]) and approaches (e.g. neural paraphrase generation [26], [28], [27]) can be used to generate paraphrases. A common problem with the generated paraphrases is that they often contain inappropriate or low-quality candidates. Hence, a paraphrase scoring module is usually employed in the downstream tasks.

In this paper, we employ a question paraphrase generation toolkit used in Baidu Search. The major advantage of this toolkit is that it can generate high-quality question paraphrases. Given 100 sampled questions, the toolkit can generate around 10 paraphrases for each question on average. We further manually evaluate the generated question paraphrases, the accuracy is around 98%. Basically, there are two modules in the toolkit.

Paraphrase Candidates Retrieval: Given a question, this module will retrieve similar questions as candidates from an inverted index of search logs in Baidu Search.

Paraphrase Similarity Model: Given each pair of a question and its retrieved paraphrase candidate, this module will estimate their semantic similarity and determines if they are paraphrases of each other based on a predefined threshold. The similarity model is a BERT-based model that was fine-tuned a large-scale labeled datasets containing more than 500K question pairs.

This toolkit benefits from both the real search engine logs and a well-tuned BERT-based similarity model. Hence, it can generate high-quality paraphrase candidates. We create a large-scale MRC dataset with the high-quality question paraphrases generated by the toolkit (see the next Section).

C. A Chinese MRC Dataset with Question Paraphrases

In this paper, we construct a large-scale Chinese MRC dataset. The way we construct this dataset is similar to SQuAD [1]. We first collect passages from Baidu Baike [1]. Then, we ask crowd sourcing workers to ask questions and annotate the corresponding answers to each passage.

Datasets	#Passages	#Questions	avg. #Que$_{paraphrases}$
Train	77.2K	217.3K	4.7
Dev	4.3K	12.1K	9.2
Test	4.3K	12.1K	9.2

Table II
THE STATISTICS OF THE MRC DATASET.

Question Class	Percentage
Entity	66.1%
Number	23.9%
Description	9.8%
Other	0.2%

Table III
THE FREQUENCY DISTRIBUTION OF DIFFERENT QUESTION CLASSES.

There are $85K$ passages and $242K$ questions in total. Additionally, the average length of questions and passages are 8.8 and 151.7 Chinese characters, respectively. The average length of answers is 6.4 characters. Table III shows the frequency distribution of different question classes. We can observe that the most frequent questions are factoid questions. Hence, the average length of answers is relatively short. As shown in Table II, we randomly divide the dataset into training, development and test sets.

We further use the question paraphrase generation toolkit to generate paraphrases for each question in each sample. On average, there are 4.7 question paraphrases for each question in training set, and 9.2 question paraphrases for each question in the development and test sets. Note that we limit the maximal number of question paraphrases in the training set to reduce the training time.

III. Approach

In this section, we first analyze the oversensitivity issue of neural MRC models. Then, we propose a regularized approach to improve the robustness of the neural models.

A. The Oversensitivity of a BERT-based Model

In this paper, we use different prediction ratio (DPR) to measure the *oversensitivity* of a neural MRC model. We define the DPR of a neural MRC model $f(\theta)$ on a dataset D as follows.

$$DPR_D(f(\theta)) = \frac{\|Q\| - \Gamma(f(\theta))}{\|Q\|},$$

where Q represents the set of original questions in dataset D, and $\|Q\|$ represents the total number of original questions without paraphrasing. Γ indicates the number of original questions whose all paraphrases make the neural model $f(\theta)$ predict the same answers. Formally, Γ is further defined as follows.

$$\Gamma = \sum_{q \in Q} \prod_{q_k \in \tilde{q}} \mathbb{1}[f(\theta; q_k, p) = f(\theta; q, p)],$$

where q_k is a paraphrase of the original question q, $\tilde{q}$ indicates the set of all the paraphrases of q, and p is a

[1] http://baike.baidu.com

Models	DEV DPR%	TEST DPR%
BERT Baseline	28.89	29.46
+ Question Paraphrases	18.93	19.63
+ Question Paraphrases, + Regularization	18.57	19.13

Table IV
THE OVERSENSITIVITY OF THE MRC MODELS.

Models	DEV		TEST	
	Rouge-L%	EM%	Rouge-L%	EM%
BERT Baseline	89.47	78.67	89.84	78.44
+ Question Paraphrases	90.28	79.57	90.54	79.33
+ Question Paraphrases, + Regularization	90.38	79.90	90.84	79.90

Table V
THE HELD-OUT ACCURACY OF THE MRC MODELS.

passage. A low DPR score means the MRC model is robust with respect to question paraphrases.

We train a BERT-based MRC model without question paraphrasing, and estimate the DPR (i.e. *oversensitivity*) of this model on both dev set and test set that contain paraphrases. The DPRs are 28.89% and 29.46% on dev set and test set, respectively. We can see that the strong BERT-based model is very sensitive to the similar inputs that are semantically equivalent.

B. A Regularized BERT-based Model

In the previous section, we can observe that the BERT-based model is not robust. To address the *oversensitivity* issue, we propose a regularized BERT-based model. Intuitively, if two questions are paraphrases of each other, a robust model should give the same answer. That is to say, the probability distributions of the answer start or end positions for question paraphrases should be closed to each other. Based on this intuition, we introduce a regularization loss by leveraging the question paraphrases, to encourage the model give closed predictions to question paraphrases. The regularization loss is defined as follows.

$$\mathcal{L}_2 = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{2(K+1)^2}\sum_{k_1,k_2=1}^{k_1,k_2=K+1}$$
$$cross\_entropy(\mathbf{p}^1_{j,k_1},\mathbf{p}^1_{j,k_2}) + cross\_entropy(\mathbf{p}^2_{j,k_1},\mathbf{p}^2_{j,k_2}),$$

where N is the total number of the original questions. K is the number of paraphrases of the j-th question. $\mathbf{p}^1_{j,k_1}$ is the probability distribution of the answer start position for the k_1-th paraphrase of j-th question. Similarly, $\mathbf{p}^2_{j,k_1}$ is the probability distribution of the answer end position for the k_1-th paraphrase of j-th question. If the neural model is robust, the regularization loss $\mathcal{L}_2$ should be small. Otherwise, the model might be *oversensitivity*.

Finally, by combining the answer prediction loss $\mathcal{L}_1$ (described in Section II-A) and the regularization loss $\mathcal{L}_2$, we expect that the model can simultaneously achieve high held-out accuracy through minimizing the answer prediction loss, and obtain high robustness through minimizing the regularization loss. Formally, the combined objective will be as follows.

$$\mathcal{L} = \mathcal{L}_1 + \lambda\mathcal{L}_2,$$

where λ is a hyper-parameter for the linear combination. In the experiment, we set λ as 1.

IV. EXPERIMENTS

In this section, we first introduce the evaluation metrics. Then, we give the comparison settings of different models and the training details. Last, we present the experimental results.

A. Evaluation Metrics

To evaluate the held-out accuracy of an MRC model, we use the following two metrics, i.e. ROUGE-L [29] and exact match (EM). ROUGE-L can be viewed as a metric to measure the partial correctness of the predicted answers. To calculate these two metrics, we first normalize the predicted and reference answers by removing spaces and punctuations. We then do the calculation in Chinese character-level.

To evaluate the *oversensitivity* of an MRC model, we use different prediction ratio (DPR) that is defined in Section III-A. The lower the DPR is, the more robust the model is.

B. Comparison Settings

In this section, we compare three models: (i) a BERT-based MRC model, that is trained on the dataset without paraphrasing, (ii) a BERT-based MRC model, that is trained on the dataset augmented by question paraphrases, (iii) a regularized BERT-based model, that is trained on the dataset augmented by question paraphrases.

C. Training Details

For all the settings of BERT-based models, we initialize parameters of the BERT encoding layer with pre-trained models officially released by Google [2]. These models were pre-trained on the corpus of Chinese Wikipedia pages, using the tasks of masked language model and next sentence prediction [13]. The pre-trained model contains 12 Transformer encoding blocks, each with 12 self-attention heads and 768 hidden units. There are $110M$

[2] https://github.com/google-research/bert

parameters in the model. Throughout our experiments, we use this setting unless specified otherwise. Other trainable parameters are randomly initialized.

We use the Adam optimizer [30] with a learning rate of 3e-5 and a batch size of 32. The number of training epochs is 2, according to the best EM and Rouge-L scores on the dev set. During training, the pre-trained BERT parameters will be fine-tuned with other trainable parameters.

D. Experimental Results

The main experimental results have been shown in Table IV and Table V.

Table IV shows the *oversensitivity* of the MRC models. We can observe that the BERT baseline is very sensitive to the question paraphrases that are semantically equivalent. By directly training the model on the dataset augmented by question paraphrases, the robustness of the model has been significantly improved. The regularized BERT-based model obtains the best robustness by leveraging the question paraphrases to encourage the model give closed predictions to question paraphrases. Comparing to the BERT baseline, the DPR decreases more than 10%.

Table V shows the held-out accuracy of the MRC models. We can observe that the BERT baseline obtains good performance on the MRC dataset. Our proposed regularized BERT-based model shows the best performance in terms of Rouge-L and EM.

In a summary, the experimental results show that our approach can significantly improve the robustness of a strong BERT-based MRC model and achieve improvements over the BERT baseline in terms of Rouge-L and EM through optimizing both the regularization loss and answer prediction loss.

V. Related Work

A. Neural Machine Reading Comprehension

In recent years, a number of large-scale datasets have been created for MRC, e.g., CNN/DM [6], SQuAD [1], SearchQA [4], TriviaQA [5], MS-MARCO [2], and DuReader [3].

These datasets have led to the advanced neural MRC models like Match-LSTM [7], BiDAF [8], DCN [9], R-Net [10], and QANet [11]. These end-to-end neural models have similar architectures, including an encoding layer, an attention-based interaction layer and a prediction layer.

Recently, the pre-trained language models such as ELMo [12], BERT [13], ERNIE [14] and XL-NET [15] have been proposed. These language models are deep neural networks, that are pre-trained on large-scale unlabeled text corpus to obtain contextual representations of text. When used in downstream tasks including MRC, the pre-trained contextual representations greatly improve the performance.

Although these neural MRC models achieves high held-out accuracy on particular datasets, they are often not robust: different ways of phrasing the same question can often cause different answers. In this paper, we focus on addressing the *oversensitivity* issue of MRC models.

B. Paraphrasing for Question Answering

A number of previous work has investigated the use of paraphrases to machine reading comprehension or question answering systems. The previous work can be classified into three categories. The first one uses question paraphrases and encourages the models to learn similar representations for the question paraphrases ([18], [19]). The second category of research directly incorporates generated question paraphrases to a question answering module by scoring them, because the generated paraphrases often contain inappropriate candidates ([17], [21], [22], [23], [24]). The third category of research mines high quality semantically equivalent adversarial rules to generate question paraphrases by involving human-in-the-loop [16].

Although the previous work tried to incorporate question paraphrases to improve the performance of the question answering systems, they did not explicitly address the *oversensitivity* issue. It is not clear to what extent the issue was addressed.

In contrast, we quantitively analyze the *oversensitivity* issue of a BERT-based MRC model and propose a regularized BERT-based model to improve the robustness of the model.

VI. Conclusion

In this paper, we focus on addressing the *oversensitivity* issue of neural machine reading comprehension (MRC) models. To address this issue, we first create a large-scale Chinese MRC dataset with high-quality question paraphrases generated by a toolkit used in Baidu Search. Then, we quantitively analyze the *oversensitivity* issue of the neural MRC models. Intuitively, if two questions are paraphrases of each other, a robust model should give the same predictions. Based on this intuition, we propose a regularized BERT-based model to encourage the model give the same predictions to similar inputs by leveraging high-quality question paraphrases. The experimental results show that our approach can significantly improve the robustness of a strong BERT-based MRC model and achieve improvements over the BERT baseline in terms of Rouge-L and EM through optimizing both the regularization loss and answer prediction loss.

In the future work, we will investigate this idea with adversarial training to further improve both the robustness and held-out accuracy of the neural MRC models.

References

[1] Rajpurkar P, Zhang J, Lopyrev K, Liang P. *Squad: 100,000+ questions for machine comprehension of text.* In Proceedings of EMNLP. 2016.

[2] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. *Ms marco: A human generated machine reading comprehension dataset.* In Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches at NIPS. 2017.

[3] W He, K Liu, J Liu, Y Lyu, S Zhao, X Xiao, Y Liu, Y Wang, H Wu, Q She. *DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications*. In Proceedings of Machine Reading for Question Answering (MRQA) Workshop at ACL. 2018.

[4] Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. *Searchqa: A new q&a dataset augmented with context from a search engine.* arXiv:1704.05179. 2017.

[5] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. *Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension.* In Proceedings of ACL. 2017.

[6] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt,Will Kay, Mustafa Suleyman, and Phil Blunsom. *Teaching machines to read and comprehend.* In Proceedings of NIPS. 2015.

[7] Shuohang Wang and Jing Jiang. *Machine comprehension using match-lstm and answer pointer.* In Proceedings of ICLR. 2017.

[8] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. *Bidirectional attention flow for machine comprehension.* In Proceedings of ICLR. 2017.

[9] Caiming Xiong, Victor Zhong, and Richard Socher. *Dynamic coattention networks for question answering.* In Proceedings of ICLR. 2017.

[10] Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. *Gated self-matching networks for reading comprehension and question answering.* In Proceedings of ACL. 2017.

[11] AdamsWei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. *Qanet: Combining local convolution with global self-attention for reading comprehension.* In Proceedings of ICLR. 2018.

[12] Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. *Deep contextualized word representations.* In Proceedings of NAACL. 2018.

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

[14] Sun Y, Wang S, Li Y, Feng S, Chen X, Zhang H, Tian X, Zhu D, Tian H, Wu H. *ERNIE: Enhanced Representation through Knowledge Integration.* arXiv preprint arXiv:1904.09223. 2019.

[15] Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov R, Le QV. *XLNet: Generalized Autoregressive Pretraining for Language Understanding.* arXiv preprint arXiv:1906.08237. 2019.

[16] Ribeiro MT, Singh S, Guestrin C. *Semantically equivalent adversarial rules for debugging nlp models.* In Proceedings of ACL. 2018.

[17] Jonathan Berant and Percy Liang. *Semantic parsing via paraphrasing.* In Proceedings of ACL. 2014.

[18] Antoine Bordes, Sumit Chopra, and Jason Weston. *Question answering with subgraph embeddings.* In Proceedings of EMNLP. 2014.

[19] Li Dong, Furu Wei, Ming Zhou, and Ke Xu. *Question answering over freebase with multicolumn convolutional neural networks.* In Proceedings of ACL. 2015.

[20] Pablo Duboue and Jennifer Chu-Carroll. *Answering the question you wish they had asked: The impact of paraphrasing for question answering.* In Proceedings of NAACL. 2006.

[21] Shashi Narayan, Siva Reddy, and Shay B Cohen. *Paraphrase generation from Latent-Variable PCFGs for semantic parsing.* In Proceedings of NLG. 2016.

[22] Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. *Paraphrase-driven learning for open question answering.* In Proceedings of ACL. 2013.

[23] Bo Chen, Le Sun, Xianpei Han, and Bo An. *Sentence rewriting for semantic parsing.* In Proceedings of ACL. 2016.

[24] Dong L, Mallinson J, Reddy S, Lapata M. *Learning to paraphrase for question answering*. In Proceedings of ACL. 2017.

[25] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. *PPDB: the paraphrase database.* In Proceedings of NAACL. 2013.

[26] Hasan SA, Lee K, Datla V, Qadir A, Liu J, Farri O. *Neural Paraphrase Generation with Stacked Residual LSTM Networks.* In Proceedings of COLING. 2016.

[27] Li Z, Jiang X, Shang L, Li H. *Paraphrase Generation with Deep Reinforcement Learning.* In Proceedings of EMNLP. 2018.

[28] Mallinson J, Sennrich R, Lapata M. *Paraphrasing revisited with neural machine translation.* In Proceedings of EACL. 2017.

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

[30] Diederik P. Kingma and Jimmy Ba. *Adam: A method for stochastic optimization..* CoRR,abs/1412.6980. 2014.