the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. Bleu might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71.

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Lila R Gleitman and Henry Gleitman. 1970. Phrase and paraphrase: Some innovative uses of language.

Tanya Goyal and Greg Durrett. 2020. Neural syntactic preordering for controlled paraphrase generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 238–252.

Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. 2021. Fast and accurate neural machine translation with translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3170–3180.

Chaitra Hegde and Shrikumar Patil. 2020. Unsupervised paraphrase generation using pre-trained language models. *arXiv preprint arXiv:2006.05477*.

Jonathan Herzig and Jonathan Berant. 2019. Don't paraphrase, detect! rapid and effective data collection for semantic parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3810–3820.

Kuan-Hao Huang and Kai-Wei Chang. 2021. Generating syntactically controlled paraphrases without using annotated parallel pairs. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1022–1033.

Tomoyuki Kajiwara. 2019. Negative lexically constrained decoding for paraphrase generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6047–6052.

Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar. 2020. Syntax-guided controlled generation of paraphrases. *Transactions of the Association for Computational Linguistics*, 8:330–345.

Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. 2019. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. 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 3609–3619.

Wuwei Lan and Wei Xu. 2018. Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3890–3902.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Xianggen Liu, Lili Mou, Fandong Meng, Hao Zhou, Jie Zhou, and Sen Song. 2020. Unsupervised paraphrasing by simulated annealing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 302–312.

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 preprint arXiv:1907.11692*.

Nitin Madnani, Joel Tetreault, and Martin Chodorow. 2012. Re-examining machine translation metrics for paraphrase identification. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 182–190.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Tong Niu, Semih Yavuz, Yingbo Zhou, Nitish Shirish Keskar, Huan Wang, and Caiming Xiong. 2021. Unsupervised paraphrasing with pretrained language models. In *Proceedings of the 2021 Conference on*

Empirical Methods in Natural Language Processing, pages 5136–5150.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André FT Martins, and Alon Lavie. 2021. Are references really needed? unbabel-ist 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040.

Raphael Shu, Hideki Nakayama, and Kyunghyun Cho. 2019. Generating diverse translations with sentence codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1823–1827.

AB Siddique, Samet Oymak, and Vagelis Hristidis. 2020. Unsupervised paraphrasing via deep reinforcement learning. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1800–1809.

Jiao Sun, Xuezhe Ma, and Nanyun Peng. 2021. Aesop: Paraphrase generation with adaptive syntactic control. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5176–5189.

Clifford H Wagner. 1982. Simpson's paradox in real life. *The American Statistician*, 36(1):46–48.

Shan Wu, Bo Chen, Chunlei Xin, Xianpei Han, Le Sun, Weipeng Zhang, Jiansong Chen, Fan Yang, and Xunliang Cai. 2021. From paraphrasing to semantic parsing: Unsupervised semantic parsing via synchronous semantic decoding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5110–5121.

Wei Xu, Chris Callison-Burch, and William B Dolan. 2015. Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 1–11.

Wei Xu, Alan Ritter, Chris Callison-Burch, William B Dolan, and Yangfeng Ji. 2014. Extracting lexically divergent paraphrases from twitter. *Transactions of the Association for Computational Linguistics*, 2:435–448.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086.

A Details of Twitter-Para

Our Twitter-Para is a pre-processed dataset based on (Xu et al., 2014, 2015). In the original dataset (Xu et al., 2014, 2015), there are some input sentences that have no corresponding references, so we drop such input-candidate pairs to create Twitter-Para. Specifically, the human-annotated score ranges from 0$\sim$1.0, where higher scores mean better quality. The basic statistics of Twitter-Para are listed in Table 11.

#input	#candidate	#reference	avg candidate
761	7159	761	9.41

Table 11: The statistics of Twitter-Para. There are 761 input sentences and each input sentence corresponds to one standard reference. Besides, there are 7159 paraphrase candidates totally, and each input sentence owns 9.41 paraphrase candidates averagely.

B Details of BQ-Para

Considering the absence of Chinese paraphrase evaluation benchmarks, we build BQ-Para based on the BQ dataset. We select 550 sentences as input sentences from BQ-dataset. Each sentence owns a manually-written reference and also owns ten candidates. Specifically, such candidates are generated by popular paraphrase generation algorithms. Then, for such a candidate, given the input sentence, we hire professional annotators to provide a score between $0 - 1.0$ to reflect its paraphrase quality. The basic statistics of BQ-Para are listed in Table 12.

#input	#candidate	#reference	avg candidate
550	5550	550	10

Table 12: The statistics of BQ-Para. There are 550 input sentences and each input sentence corresponds to one standard reference. Besides, there are 5550 paraphrase candidates totally, and each input sentence owns 10 paraphrase candidates averagely.

C Definition of normalized edit distance

Given two sentences $\mathbf{x}$ and $\mathbf{x}^i$, the definition of normalized edit score is defined as follows:

$$NED = \frac{\text{dist}\left(\mathbf{x}, \mathbf{x}^i\right)}{\max\left(|\mathbf{x}|, |\mathbf{x}^i|\right)} \tag{9}$$

where $|\mathbf{x}|$ is the length of sentence $\mathbf{x}$.

D Definition of BERT-iBLEU and iBLEU

BERT-iBLEU is defined as follows:

$$\begin{aligned} \text{BERT-iBLEU} &= \frac{\beta + 1.0}{\beta \cdot \text{BERTScore}^{-1} + 1.0 \cdot (1 - \text{SelfBLEU})^{-1}} \\ \text{SelfBLEU} &= \text{BLEU}(\text{input, candidate}) \end{aligned} \tag{10}$$

where β is a constant (usually set as 4).

iBLEU is a hybrid metric that computes the difference between BLEU and SelfBLEU, which is defined as follows:

$$\text{iBLEU} = \text{BLEU} - \alpha \cdot \text{SelfBLEU} \tag{11}$$

where α is a constant (usually set as 0.3).

E A detailed analysis towards BERT-iBLEU

Principally, we can formulate any existing metrics into the combination of semantic similarity (Sim) and lexical divergence(Div), including BERT-iBLEU. Firstly, we recall the definition of BERT-iBLEU:

$$\text{BERT-iBLEU} = \frac{\beta + 1.0}{\beta \cdot \text{BERTScore}^{-1} + 1.0 \cdot (1 - \text{SelfBLEU})^{-1}}$$

Naturally, we re-write BERT-iBLEU as the following formation:

$$\text{BERT-iBLEU} = \frac{\beta + 1.0}{\beta \cdot \text{Sim}^{-1} + \cdot (\text{Div})^{-1}}$$

where Sim represents the BERTScore and Div denotes (1-SelfBLEU). Though such a formation indeed contains both lexical divergence and semantic similarity, it can not gaurantee that BERT-iBLEU is a good paraphrase metric that serves as a human-like automatic metric. Existing work (Niu et al., 2021) only shows that it outperforms n-gram-based metrics. The following experiments demonstrate an interesting conclusion: *BERT-iBLEU consistently performs worse than SelfBERTScore*, and then we present our analysis. The results are demonstrated in Table 13, from where we can see that BERT-iBLEU(B) consistently under-perform than BERTScore(B).

Metric	Twitter-Para		BQ-Para	
	Pr.	Spr.	Pr.	Spr.
BERTScore(B).Free	0.491	0.488	0.397	0.392
BERT-iBLEU(B,4)	0.488	0.485	0.393	0.383
BERT-iBLEU(B,5)	0.490	0.488	0.395	0.392
BERT-iBLEU(B,10)	0.490	0.488	0.396	0.389

Table 13: The Pearson (Pr.) and Spearman (Spr.) correlations of vanilla BERTScore and BERT-iBLEU. We can see BERT-iBLEU consistently under-perform vanilla BERTScore on both benchmarks.

To explain such interesting results, we re-write BERT-iBLEU as follows:

$$\begin{aligned} \text{BERT-iBLEU} &= \frac{\beta + 1.0}{\beta \cdot \text{Sim}^{-1} + \cdot (\text{Div})^{-1}} \\ &= \frac{\beta \cdot \text{Sim} \cdot \text{Div} + \text{Sim} \cdot \text{Div}}{\beta \cdot \text{Div} + \text{Sim}} \\ &= \text{Sim} + \frac{\text{Sim} \cdot \text{Div} - \text{Sim}^2}{\beta \cdot \text{Div} + \text{Sim}} \end{aligned}$$

As we can see, BERT-iBLEU can be decoupled into two terms Sim and $\frac{\text{Sim}\cdot\text{Div}-\text{Sim}^2}{\beta\cdot\text{Div}+\text{Sim}}$ (We denote it as term 'Mix'). According to the analysis in our paper, after removing the Sim, the remaining part, the 'Mix' term should be able to reflect diversity. However, the 'Mix' term does not represent meaningful aspects of paraphrase quality. Specifically, we investigate the correlation between the 'Mix' term and human annotation, only resulting in correlations close to zero, indicating that the 'Mix' term is improper since there is nearly no correlation between it and human annotation. Overall, BERT-iBLEU owns an improper combination of semantic similarity and diversity.