Metric	Twitter-Para	BQ-Para
BLEU-4.Free	-0.197	-0.075
Rouge-1.Free	-0.385	-0.334
Rouge-2.Free	-0.377	-0.308
Rouge-L.Free	-0.426	-0.514
METEOR.Free	-0.233	-
BERTScore(B).Free	-0.424	-0.347
BARTScore.Free	-0.187	-0.263
NED	0.635	0.655

Table 8: Pearson correlation of ΔM and Δh on $\mathcal{S}_{\text{div1}}$.

5 New Metric: ParaScore

5.1 ParaScore

Inspired by previous experiments and analyses, we propose a new metric named ParaScore, as below,

$$\text{ParaScore} = \max(Sim(X,C), Sim(R,C)) + \omega \cdot DS(X,C) \quad (7)$$

where ω is a hyper-parameter in our experiments, $\max(Sim(X,C), Sim(R,C))$ is motivated by the analysis in §3.2, and DS is defined as a sectional function to model lexical divergence (referring to the analysis in §4.3):

$$DS(X,C) = \begin{cases} \gamma & \text{d} > \gamma \\ \text{d} \cdot \frac{\gamma+1}{\gamma} - 1 & 0 \le \text{d} \le \gamma \end{cases} \quad (8)$$

where γ is a hyper-parameter, d=$Dist(X,C)$, which can be any proper distance metric. In our experiments, Sim and Div are respectively instantiated by BERTScore and NED [6], and γ is fixed as 0.35.

ParaScore defined in Eq. (7) involves the reference R and thus it is a reference-based metric. It is natural to extend ParaScore to its reference-free version **ParaScore.Free** by removing R as follows:

$$\text{ParaScore.Free} = Sim(X,C) + \omega \cdot DS(X,C).$$

5.2 Experimental Results

Benchmarks and baselines Experiments are conducted on four datasets: Twitter-Para, BQ-Para, and the extended version of them. The extended version of each dataset is built by adding 20% of the input sentences as candidates. They are called **Twitter(Extend)** and **BQ(Extend)** respectively. Since the newly added candidates are input sentences, according to the requirements of paraphrasing, their annotation scores are 0. The goal of adding the extended version of the datasets is to test the robustness of different metrics on various data distributions. In addition to the baselines in previous sections, we add two more baselines: **BERT-iBLEU** (Niu et al., 2021) and **iBLEU** (Siddique et al., 2020; Liu et al., 2020), whose details are listed in Appendix D.

[6]Note that there may be other advanced metrics to instantiate Sim (e.g., SimCSE) and other heuristic combination (e.g., weighted geometric mean) methods, which we leave as future work.

Metric	Twitter-Para		BQ-Para	
	Pearson	Spearman	Pearson	Spearman
BERTScore(B)	0.470	0.468	0.332	0.322
BERTScore(R)	0.368	0.358	0.387	0.376
BARTScore	0.311	0.306	0.260	0.246
iBLEU(0.2)	0.013	0.033	0.155	0.139
BERTScore(B).Free	0.491	0.488	0.397	0.392
BERT-iBLEU(B,4)	0.488	0.485	0.393	0.383
ParaScore	**0.522**	**0.523**	**0.492**	**0.489**
ParaScore.Free	0.492	0.489	0.398	0.393
Metric	Twitter(Extend)		BQ-Para(Extend)	
	Pearson	Spearman	Pearson	Spearman
BERTScore(B)	0.427	0.432	0.248	0.267
BERTScore(R)	0.334	0.329	0.299	0.317
BARTScore	0.280	0.276	0.199	0.206
iBLEU(0.2)	0.011	0.032	0.129	0.121
BERTScore(B).Free	0.316	0.419	0.230	0.312
BERT-iBLEU(B,4)	0.327	0.416	0.221	0.303
ParaScore	**0.527**	**0.530**	**0.510**	**0.442**
ParaScore.Free	0.496	0.495	0.487	0.428

Table 9: The Pearson (Pr.) and Spearman (Spr.) correlations on two benchmarks. Specifically, we highlight the best performance with **Bold numbers**. BERT-iBLEU(B,4) means the encoder is BERT and β is 4. iBLEU(0.2) indicates α is set as 0.2.

Performance comparison The performance of each metric on the four datasets are listed in Table 9. Several observations can be made. First of all, ParaScore performs significantly better than all the other metrics on all the datasets. It is also shown that ParaScore is much more robust than other metrics. Second, on both Twitter-Para and BQ-Para, BERT-iBLEU performs worse than vanilla BERTScore. Note that BERT-iBLEU (Niu et al., 2021) also considers lexical divergence, and it applies a harmonic weight mean of BERTScore (for semantic similarity) and -BLEU.Free (for lexical divergence). However, according to results in Table 9, it is only comparable to BERTScore.Free or even worse. This further indicates that 1) the weighted harmonic mean formation is sub-optimal, 2) the sectional threshold is important as discussed in §4.3, making the performance comparable to

BERTScore.Free in most cases, as shown in Appendix E.

Metric	Twitter(Extend)		BQ-Para(Extend)	
	Pr.	Spr.	Pr.	Spr.
ParaScore	**0.527**	**0.530**	**0.510**	**0.442**
ParaScore w/o thresh	0.358	0.450	0.266	0.333
ParaScore w/o max	0.496	0.495	0.487	0.428
ParaScore w/o DS	0.349	0.450	0.249	0.326

Table 10: Ablation study on the ParaScore. ParaScore w/o thresh means removing the sectional formation defined in Eq 8. ParaScore w/o DS means removing the lexical divergencescore.

Ablation study We study of effect of three factors of ParaScore: the max function, the DS function for divergence, and the threshold mechanism in Equ (8). The results are listed in Table 10. By comparing ParaScore with 'ParaScore w/o DS', we can see that ParaScore significantly degrades when removing DS or its sectional version, which confirms the effectiveness of DS and the sectional function for DS. These findings demonstrate that a sectional function for Div is beneficial for paraphrase evaluation. According to the results, all of the above listed factors are essential for the effectiveness of ParaScore.

Discussion According to Table 8, we can observe that existing metrics do not well consider the lexical divergence, including BERTScore.Free. On the two original benchmarks, as shown in Table 9, BERTScore.Free is still competitive with ParaScore.Free, which explicitly models lexical divergence. This fact seems to disagree with the human evaluation guideline that lexical divergence is also important. Therefore, these results may reveal a potential drawback in the original benchmarks: They overlook the role of lexical divergence. Although the extended version of both benchmarks alleviates such a drawback to some extent, it introduces divergence into both datasets in a toy manner by copying the inputs rather than in a natural manner. It would be important to build a better benchmark for paraphrase evaluation in the future.

6 Related Work

Most previous works conduct paraphrase evaluation by the reference-based MT metrics from the popular tasks similar to paraphrase generation such as machine translation (Bannard and Callison-Burch, 2005; Callison-Burch, 2008; Cohn et al., 2008; Kumar et al., 2020; Goyal and Durrett, 2020; Sun et al., 2021; Huang and Chang, 2021). However, paraphrase evaluation is different from these tasks: the paraphrase should possess lexical or syntactic differences toward the input sentence, which is not emphasized in these tasks.

Generally, the metrics in paraphrase evaluation can be divided into two kinds: reference-free and reference-based metric. Most reference-based metrics include BLEU (Papineni et al., 2002), Rouge (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). In addition, the reference-free of these metrics have also been used: Self-BLEU (Shu et al., 2019) measures the BLEU score between the generated paraphrase and input sentence. Moreover, the iBLEU (Choshen and Abend, 2018) score penalizes repeating the input sentence in the generated paraphrase. BERT-iBLEU (Zhou and Bhat, 2021) takes the weighted harmonic mean of the BERTscore (Zhang et al., 2019) and one minus self-BLEU. Previous works commonly utilize reference-based metrics in evaluation, in this paper, we also pay attention to the overlooked reference-free metrics.

The difference between the existing works and our work is obvious. Existing works mainly employ these metrics to evaluate the paraphrases generated from a model. However, the reliability of existing paraphrase metrics has not been evaluated comprehensively. Thus, we prepare two paraphrase evaluation benchmarks (Chinese and English) and conduct comprehensive experiments to compare existing metrics' performance on these benchmarks. In particular, based on the empirical findings, this paper proposes a new framework for paraphrase evaluation.

7 Conclusion

This paper first reviews the reliability of existing metrics for paraphrasing evaluation by investigating how well they correlate with human judgment. Then, we find two interesting findings and further ask two questions behind them that are overlooked by the community: (1) why do reference-free metrics outperform reference-based ones? (2) what is the limitation of existing metrics? We deliver detailed analyses of such two questions and present the explanation by disentangling paraphrase quality. Based on our analyses, finally, we propose ParaScore (with both reference-based and reference-free implementations) for paraphrase evaluation, and

its effectiveness is validated through comprehensive experiments. In addition, we call for building better benchmarks which can faithfully reflect the importance of lexical divergence in paraphrase evaluation; we hope it will shed light on the future direction.

Limitation

One limitation in this paper is that it does not provide a perfect benchmark which remarkably reflects the importance of lexical divergence in a natural way rather than the heuristic way used in the experiments. Creating such a benchmark would be important for future studies on paraphrase evaluation. It is also interesting to examine the potential benefits of the proposed ParaScore on such a benchmark.

Ethical Considerations

The datasets used in this paper will not pose ethical problems. For the Twitter-Para dataset, it is a publicly available dataset. For the BQ-Para dataset, its inputs are from the public dataset BQ and we recruited five annotators to manually annotate the quality of paraphrases with the proper pay.

References

Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. 2019. Comqa: A community-sourced dataset for complex factoid question answering with paraphrase clusters. 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 307–317.

Icek Ajzen and Martin Fishbein. 1975. A bayesian analysis of attribution processes. *Psychological bulletin*, 82(2):261.

W Thomas Anderson Jr, Eli P Cox III, and David G Fulcher. 1976. Bank selection decisions and market segmentation: Determinant attribute analysis reveals convenience-and sevice-oriented bank customers. *Journal of marketing*, 40(1):40–45.

Marianna Apidianaki, Guillaume Wisniewski, Anne Cocos, and Chris Callison-Burch. 2018. Automated paraphrase lattice creation for hyter machine translation evaluation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 480–485.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Deng Cai, Yan Wang, Huayang Li, Wai Lam, and Lemao Liu. 2021. Neural machine translation with monolingual 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 7307–7318.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 196–205.

Ruisheng Cao, Su Zhu, Chenyu Yang, Chen Liu, Rao Ma, Yanbin Zhao, Lu Chen, and Kai Yu. 2020. Unsupervised dual paraphrasing for two-stage semantic parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6806–6817.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Jing Chen, Qingcai Chen, Xin Liu, Haijun Yang, Daohe Lu, and Buzhou Tang. 2018. The bq corpus: A large-scale domain-specific chinese corpus for sentence semantic equivalence identification. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4946–4951.

Leshem Choshen and Omri Abend. 2018. Automatic metric validation for grammatical error correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1372–1382.

Trevor Cohn, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics*, 34(4):597–614.

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*