On the Evaluation Metrics for Paraphrase Generation

Lingfeng Shen
Johns Hopkins University

Lemao Liu
Tencent AI Lab

Haiyun Jiang
Tencent AI Lab

Shuming Shi
Tenent AI Lab

Abstract

In this paper we revisit automatic metrics for paraphrase evaluation and obtain two findings that disobey conventional wisdom: (1) Reference-free metrics achieve better performance than their reference-based counterparts. (2) Most commonly used metrics do not align well with human annotation. Underlying reasons behind the above findings are explored through additional experiments and in-depth analyses. Based on the experiments and analyses, we propose ParaScore, a new evaluation metric for paraphrase generation. It possesses the merits of reference-based and reference-free metrics and explicitly models lexical divergence. Experimental results demonstrate that ParaScore significantly outperforms existing metrics. The codes and toolkit are released in https://github.com/shadowkiller33/ParaScore.

1 Introduction

Paraphrase generation is a fundamental problem in natural language processing (NLP), which has been widely applied in versatile tasks, such as question answering (Dong et al., 2017; Lan and Xu, 2018; Gan and Ng, 2019; Abujabal et al., 2019), machine translation (Madnani et al., 2012; Apidianaki et al., 2018; Kajiwara, 2019), and semantic parsing (Herzig and Berant, 2019; Wu et al., 2021; Cao et al., 2020). Recent years have witnessed rapid development in paraphrase generation algorithms (Sun et al., 2021; Huang and Chang, 2021; Kumar et al., 2020). However, little progress has been made in the automatic evaluation of this task. It is even unclear which metric is more reliable among many widely used metrics.

Most evaluation metrics used in previous paraphrase generation research are not designed for the task itself, but adopted from other evaluation tasks, such as machine translation (MT) and summarization. However, paraphrase evaluation is inherently different from the evaluation of most other tasks, because a good paraphrase typically obeys two criteria (Gleitman and Gleitman, 1970; Chen and Dolan, 2011; Bhagat and Hovy, 2013): semantic similarity (Sim) and lexical divergence (Div). Sim means that the paraphrase maintains similar semantics to the input sentence, whereas Div requires that the paraphrase possesses lexical or syntactic differences from the input. In contrast, tasks like machine translation have no requirement for Div. It is therefore uncertain whether the metrics borrowed from other tasks perform well in paraphrase evaluation.

In this paper, we revisit automatic metrics for paraphrase evaluation. We collect a list of popular metrics used in recent researches (Kumar et al., 2019; Feng et al., 2021; Hegde and Patil, 2020; Sun et al., 2021; Huang and Chang, 2021; Kumar et al., 2020), and computed their correlation with human annotation. Generally, these metrics fall into two categories, i.e., reference-based and reference-free metrics. The former is utilized much more frequently than the latter.

We first empirically quantify the matching degree between metric scores and human annotation, on two datasets of different languages. Upon both benchmarks, we make comprehensive experiments to validate the reliability of existing metrics. Surprisingly, we obtain two important findings: (1) Reference-free metrics better align with human judgments than reference-based metrics on our benchmarks, which is counter-intuitive in related evaluation tasks. (2) Most of these metrics (especially the commonly-used BLEU and Rouge) do not agree well with human evaluation.

Then we explore the potential reasons behind the above findings through additional experiments. For the first finding, we demonstrate that the performance comparison between reference-free and reference-based metrics is largely affected by the input-candidate and reference-candidate distance distribution. Specifically, *reference-free metrics*

are better because most paraphrase candidates in the testset have larger lexical distances to the reference than to the input, but reference-based metrics may be better for the minority candidates. To study the second finding, we design an approach based on attribution analysis (Ajzen and Fishbein, 1975; Anderson Jr et al., 1976) to decouple the factors of semantic similarity and lexical divergence. Our experiments and analysis show that *existing metrics measure semantic similarity well, but tend to ignore lexical divergence*.

Based on our analyses, we propose a new family of metrics named ParaScore for paraphrase evaluation, which takes into account the merits from both reference-based and reference-free metrics and explicitly models lexical divergence. Extensive experiments show that our proposed metrics significantly outperform the ones employed in previous research.

In summary, our main contributions are: [1]

• We observe two interesting findings that disobey conventional wisdom. First, reference-free metrics outperform reference-based ones on our benchmarks. Second, most existing metrics do not align well with human annotation.

• Underlying reasons behind the above findings are explored through additional experiments and in-depth analysis.

• Based on the findings and analysis, we propose ParaScore, a family of evaluation metrics for paraphrase generation. They align significantly better with human annotation than existing metrics.

2 Revisiting Paraphrasing Metrics

2.1 Settings

In a standard supervised paraphrase evaluation scenario, we are given an input sentence X and a reference R (the golden paraphrase of X). The goal is to evaluate the quality of a paraphrase candidate C.

Dataset Our experiments selected two benchmarks: Twitter-Para (English) and BQ-Para (Chinese). Twitter-Para is from the Twitter dataset (Xu et al., 2014, 2015), while BQ-Para is built based on the BQ dataset (Chen et al., 2018).

Specifically, considering that some metrics may have hyper-parameters, so we use 10% data in the benchmark as the dev set and tune the hyper-parameters on the dev set. Then the performance of metrics is evaluated on the remaining 90% data. Please refer to Appendix A for more details about the two datasets.

Chosen Metrics We select the following well known metrics: **BLEU** (Papineni et al., 2002), **ROUGE** (Lin, 2004), **METEOR** (Banerjee and Lavie, 2005), **BERTScore** (Zhang et al., 2019), and **BARTScore** (Yuan et al., 2021). Specifically, we consider two variants of BERTScore: **BERTScore(B)** and **BERTScore(R)**, based on BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) respectively. For each metric M, we consider its two variants, i.e., a reference-based version and a reference-free version 'M.Free'. In the reference-free version, the quality of a candidate C is estimated by $M(C, X)$, where X is the input. Similarly, in the reference-based version, the formula is $M(C, R)$, where R is the reference.

2.2 Experimental Results

Metric	Twitter-Para		BQ-Para	
	Pr.	Spr.	Pr.	Spr.
BLEU-4	-0.119	-0.104	0.127	0.144
BLEU-4.Free	-0.113↑	-0.101↑	0.109↓	0.136↓
Rouge-1	0.271	0.276	0.229	0.206
Rouge-1.Free	0.292↑	0.300↑	0.264↑	0.232↑
Rouge-2	0.181	0.144	0.226	0.216
Rouge-2.Free	0.228↑	0.189↑	0.252↑	0.242↑
Rouge-L	0.249	0.239	0.221	0.204
Rouge-L.Free	0.266↑	0.253↑	0.260↑	0.230↑
METEOR	0.423	0.418	-	-
METEOR.Free	0.469↑	0.471↑	-	-
BERTScore(B)	0.470	0.468	0.332	0.322
BERTScore(B).Free	0.491↑	0.488↑	0.397↑	0.392↑
BERTScore(R)	0.368	0.358	0.387	0.376
BERTScore(R).Free	0.373↑	0.361↑	0.449↑	0.438↑
BARTScore	0.311	0.306	0.241	0.230
BARTScore.Free	0.295↓	0.286↓	0.282↑	0.263↑

Table 1: The Pearson (Pr.) and Spearman (Spr.) correlations between popular metrics and human judgments on two datasets. **Red** text (or the text followed by '↑') indicates that reference-free metrics are better, whereas **blue** text (or the text followed by '↓') means the opposite. Please note that we do not apply METEOR to BQ-Para since METEOR is based on the English WordNet (Miller, 1995).

For each dataset and metric, the score of each sentence in the dataset is calculated by the met-

[1] The new dataset and the code of ParaScore is available at the supplementary materials.

ric. The obtained scores are then compared with human annotation to check their correlation. The correlation scores, measured by Pearson and Spearman correlations, are reported in Table 1. Several observations can be made from the table.

Reference-based vs. reference-free It can be seen from the table that, for most metrics, their reference-free variants align better with human annotation than their reference-based counterparts. This indicates that reference-free metrics perform better in the paraphrase generation task, which is somewhat counterintuitive. More detailed analysis about this observation will be given in Sec 3.

Low correlation The second observation is that, the correlation between the metrics and human judgments is not high. In other words, most of the commonly-used metrics do not align well with human annotation. BLEU-4 even shows a negative correlation with human annotation on Twitter-Para. As the third observation, embedding-based metrics (e.g., BERTScore) tend to outperform ngram-based ones (including the variants of BLEU and Rouge). The main reason for this lies in the effectiveness of embedding-based metrics in capturing semantic similarity. Despite the better performance, embedding-based metrics are still far from satisfactory. On one hand, the results in the table show that the correlation scores for the embedding-based metrics are not high enough. On the other hand, embedding-based metrics assign a very high score for a candidate if it is the same as the input text. This is an obvious flaw, because it violates the lexical divergence criterion of paraphrasing. Therefore, we can see obvious drawbacks for both ngram-based and embedding-based metrics.

In summary, we have two findings from the experimental results: (1) Reference-free metrics outperform reference-based ones on our benchmarks. (2) Most of the popular metrics (especially the commonly-used BLEU and Rouge) do not align well with human annotation.

Since the two findings are more or less surprising, some study is necessary to reveal the underlying reasons behind the findings. We hope the study helps to discover better metrics for paraphrase generation. In-depth analysis to the two findings are shown in Sec 3 and Sec 4 respectively.

3 Reference-Free vs. Reference-Based

The results in the previous section indicate that reference-free metrics typically have better performance than their reference-based counterpart. In this section, we investigate this finding by answering the following question: When and why do reference-free metrics outperform their reference-based variants?

3.1 The Distance Effect

Recall that the reference-based and reference-free variants of a metric M calculate the score of a candidate sentence C by $M(C, R)$ and $M(C, X)$ respectively. Intuitively, as shown in (Freitag et al., 2020; Rei et al., 2021), if the lexical distance between C and R is large, $M(C, R)$ may not agree well with human annotation. Therefore, we guess the lexical distance $Dist(C, R)$ between C and R may be an important factor that influences the performance of $M(C, R)$ with respect to human evaluation.

To verify this conjecture, we divide the candidates in a benchmark (e.g., Twitter-Para) into four equal-size groups (group 1 to group 4) according to $Dist(C, R)$,[2] where elements in group 1 have small $Dist(C, R)$ values. The performance of several reference-based metrics on such four groups is shown in Figure 1[3]. It can be seen that when $Dist(C, R)$ grows larger, the performance of the metrics decreases. There is a significant performance drop from group 3 to group 4 when the lexical distance is very large.

Similarly experiments are conducted for reference-free metrics. We separate the sentences in Twitter-Para into four equal-size groups according to $Dist(C, X)$ and obtain the results in Figure 2. Again, the correlation between each metric and human annotation decreases when $Dist(X, C)$ gets larger. A significant performance drop is observed when the lexical distance is very large (see group 4). The above results indicate that small lexical distances are important for both reference-based and reference-free metrics to produce high-quality scores.

[2]Here $Dist$ is measured by normalized edit distance (NED), which is widely used in retrieving translation memory (Cai et al., 2021; He et al., 2021). Its definition is deferred to Appendix C .

[3]We can see a counter-intuitive observation that the highest correlation on the subset is lower than the one on the whole set. This is a reasonable statistical phenomenon called Simpson's paradox (Wagner, 1982).