

Figure 1: Performance of reference-based metrics significantly degrades as $Dist(R, C)$ becomes large.

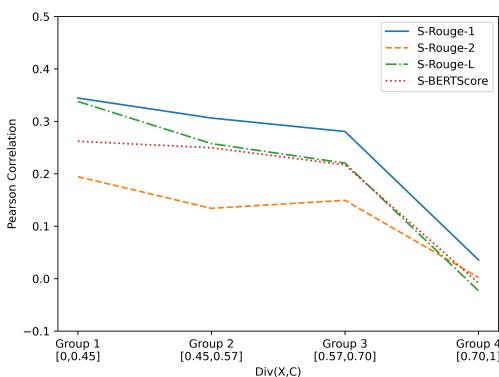


Figure 2: Performance of reference-free metrics significantly degrades as $Dist(X, C)$ becomes large.

3.2 Average Distance Hypothesis

According to the results in the previous subsection, the average distance from a group of candidates to R or X has a large effect on the performance of a metric on this candidate group. It is reasonable to further guess that lexical distances also affects the performance comparison between reference-based and reference-free metrics.

Therefore we make the following **average distance hypothesis**:

For a group of candidates G , a reference-based metric outperforms its reference-free counterpart on G if $Dist(G, R)$ is significantly larger than $Dist(G, X)$. Similarly, the reference-free version is better if $Dist(G, X)$ is greatly larger than $Dist(G, R)$.

Here $Dist(G, X)$ denotes the average lexical distance from the candidates in G to X .

To validate the above hypothesis, we divide a dataset into two parts (part-I and part-II) according to whether $Dist(C, R) > Dist(X, C)$ or not, as shown in Figure 3. Then we compare the perfor-

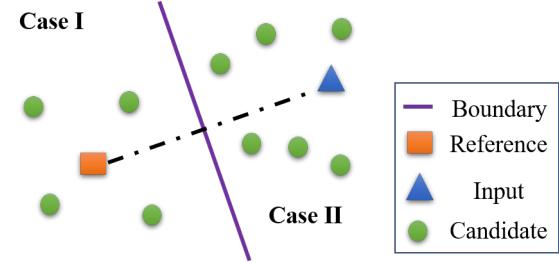


Figure 3: The boundary is the vertical parallel of the ‘reference-input’ line, which separate candidates into two cases. Case I means $Dist(R, C) < Dist(X, C)$, while Case II means the opposite.

Metric	Twitter-Para		BQ-Para	
	I	II	I	II
RougeL	0.357	0.191	0.352	0.193
RougeL.Free	0.207	0.325	0.319	0.280
Rouge1	0.367	0.223	0.362	0.199
Rouge1.Free	0.267	0.345	0.308	0.270
Rouge2	0.256	0.120	0.366	0.200
Rouge2.Free	0.160	0.275	0.283	0.260
BERTScore	0.284	0.162	0.404	0.319
BERTScore.Free	0.191	0.277	0.400	0.417
$\Delta(M.\text{Free}, M)$	-0.110	+0.132	-0.044	+0.079

Table 2: The correlation of metrics concerning human annotation on the two parts of Twitter-Para and BQ-Para. $\Delta(M.\text{Free}, M)$ denotes the averaged correlation difference between the reference-free metrics ($M.\text{Free}$) and the reference-based metrics M per metric. $\Delta(M.\text{Free}, M) > 0$ indicates the reference-free metric ($M.\text{Free}$) is better.

mance of reference-free and reference-based metrics on the two parts of data. The performance of reference-free and reference-based metrics on such cases are listed in Table 2. It is clearly shown that reference-based metrics enjoy better performance on part-I, whereas reference-free metrics perform better on part-II. Such results do verify our average distance hypothesis.

3.3 Why do Reference-Free Metrics Perform Better on our Benchmarks?

By employing the average distance hypothesis, we explain why reference-free metrics have higher performance on our datasets. We calculate the proportion of candidates in Case I and Case II (referring to Figure 3) on Para-Twitter and BQ-Twitter. The results are presented in Table 3. It is shown that there is a larger fraction of Case-II candidates than

Case-I on each dataset. Therefore, according to the average distance hypothesis, it is reasonable to have the observation that reference-free metrics are often better than their reference-based counterparts on both datasets.

Metric	Twitter-Para		BQ-Para	
	I	II	I	II
$\Delta(M.\text{Free}, M)$	-0.110	+0.132	-0.044	+0.079
Proportion	46.4%	53.6%	15.7%	84.3%

Table 3: The proportion of **Case I** and **Case II** candidates on Twitter-Para and BQ-Para. A positive $\Delta(M.\text{Free}, M)$ means reference-free metrics are better, whereas a negative value indicates that reference-based metrics have better performance.

4 Decoupling Semantic Similarity and Lexical Divergence

In this section, we investigate why most metrics do not align well with human annotation.

4.1 Attribution Analysis for Disentanglement

As illustrated earlier, a good paraphrase typically obeys two criteria: semantic similarity (*Sim*) and lexical divergence (*Div*). To seek the reasons behind the low performance of the metrics, we may need to explore how well these metrics perform in terms of each criterion. However, only one human score is available for a candidate on each dataset. The score is about the overall paraphrasing quality rather than those for a single criterion (either semantic similarity or lexical divergence).

In this section, we propose an approach to decouple the performance of the metrics in terms of each criterion. This proposed approach is inspired by attribution analysis (Anderson Jr et al., 1976; Ajzen and Fishbein, 1975) and its key idea is to analyze the attribution of one component (or dimension) while controlling the attributions from other components (or dimensions).

Applying attribution analysis to our scenario, we construct a subset $\mathcal{S} = \{(X, C_j, C_k)\}$, where (C_j, C_k) is a paraphrase candidate pair for an input sentence X , such that the difference between C_j and C_k on one criterion (*Sim* or *Div*) is significant but the difference on the other criterion is close to zero. As a result, on such a subset \mathcal{S} , the difference of human score between C_j and C_k is mainly attributed by the interested criterion. Then we can

measure the correlation between human scores and a metric in the specific criterion.

	Twitter-Para		BQ-Para	
	\mathcal{S}_{sim}	Base	\mathcal{S}_{sim}	Base
#num	583	9158	200	5156
ρ	0.805	0.345	0.629	0.394

Table 4: Pearson correlation of ΔS and Δh on \mathcal{S}_{sim} compared with that on paraphrase pairs filtered by only Eq.(1) only (Base). The results also demonstrate the necessity of the constraint Eq.(2).

Since there are no ground truth measures for *Sim* and *Div*, we use normalized edit distance (NED) and SimCSE (Gao et al., 2021) as the surrogate ground truth of *Div* and *Sim* respectively. They are chosen for two reasons. First, they are widely used and proven to be good for measuring *Div* and *Sim*. Second, they are not used as the metrics for paraphrase evaluation in this paper. Therefore, the potential unfairness is reduced.⁴

4.2 Performance in Capturing *Sim*

Formally, suppose the subset \mathcal{S}_{sim} denotes all (X, C_j, C_k) satisfying the following constraints:

$$\begin{aligned} |Dist(X, C_j) - Dist(X, C_k)| &\leq \eta_1 \\ |Sim(X, C_j) - Sim(X, C_k)| &\geq \eta_2 \end{aligned} \quad (1)$$

where *Dist* is a distance function for calculating *Div*, η_1 is set as 0.05 and η_2 is 0.15.⁵

In addition, we define two quantities for each tuple (X, C_j, C_k) from \mathcal{S}_{sim} as follows:

$$\Delta S = Sim(X, C_j) - Sim(X, C_k) \quad (2)$$

$$\Delta h = h(X, C_j) - h(X, C_k) \quad (3)$$

where *h()* refers to the human score. Then we measure the correlation between ΔS and Δh on \mathcal{S}_{sim} , and the results are shown in Table 4. It can be seen that the correlation is much higher on \mathcal{S}_{sim} compared with that on all paraphrase pairs, indicating good disentanglement on \mathcal{S}_{sim} . As \mathcal{S}_{sim} is proper to demonstrate how well a metric captures semantic similarity, we call it **semantic-promoted data**.

⁴For example, if we use BERTScore to compute *Sim*, the statistics on \mathcal{S} may be biased to BERTScore and thus becomes unfair for other metrics.

⁵Intuitively, the disentanglement effect would be better if η_1 is more close to zero and η_2 is much larger. However, this leads to the limited size of \mathcal{S}_{sim} due to the contradictory between *Sim* and *Div*, and hence the statistical correlation on \mathcal{S}_{sim} is not significant.

Metric	Twitter-Para	BQ-Para
BLEU-4.Free	0.067	0.372
Rouge-1.Free	0.574	0.430
Rouge-2.Free	0.400	0.350
Rouge-L.Free	0.481	0.388
METEOR.Free	0.499	-
BERTScore(B).Free	0.785	0.576
BARTScore.Free	0.797	0.552
Sim	0.805	0.629

Table 5: Pearson correlation of ΔM and Δh on S_{sim} , the ‘semantic-promoted data’. This is an example to show that paraphrase quality does not increase as lexical divergence increases.

To investigate how well existing metrics capture semantic similarity, we add an extra definition:

$$\Delta M = M(X, C_j) - M(X, C_k) \quad (4)$$

where M is a reference-free metric. Then we measure the correlation between ΔM and Δh on the semantic-promoted data, and get the results in Table 5. The results suggest that the embedding-based metrics (i.e., BERTScore.Free) significantly outperform word-overlap metrics (i.e., BLEU.Free) in capturing semantic similarity. Overall, the results show that some metrics perform pretty well in capturing semantic similarity.

4.3 Performance in Capturing Div

Similarly, to analyze the ability of metrics in capturing Div , we exchange $Dist$ with Sim in Eq 1 and obtain a subset of tuples named S_{div} ($\eta_1 = 0.05$ and $\eta_2 = 0.10$). In this case, the principal attribution on S_{div} is lexical divergence. In addition, we define ΔD as follows:

$$\Delta D = Dist(X, C_j) - Dist(X, C_k) \quad (5)$$

Then we conduct analyses on S_{div} to examine the effect of disentanglement for lexical divergence. It is interesting that the correlation between ΔD and Δh on S_{div} is almost zero, which indicates that higher distance scores does not guarantee better paraphrasing. This fact is in line with previous findings (Bhagat and Hovy, 2013). Let’s explain by the examples in Table 6. It is reasonable for candidate C_1 to get a low human annotation score due to its small lexical distance to the input X . Though C_3 has a larger distance to X than C_2 , they are assigned the same annotation score, possibly because both C_2 and C_3 are good enough in terms of Div from the viewpoint of human annotators.

Such results show that when the distance is large (i.e., beyond a threshold), Div does not correlate well with human score h .

Type	Text	Dist	h
X	NLP is a potential research field	-	-
C_1	NLP is a promising research field	0.21	0.4
C_2	NLP is a promising study area	0.53	1.0
C_3	The NLP field has high potential	0.79	1.0

Table 6: X and C refer to the input and candidate. This example shows that paraphrase quality annotated by human (h) does not always increase as the lexical divergence ($Dist$) increases.

We modify our decoupling strategy by further dividing S_{div} into two parts according to a distance threshold. We define d as follows:

$$d(j, k) = \min(Dist(X, C_j), Dist(X, C_k)) \quad (6)$$

where $d(j, k)$ represents the minimum $Dist$ score in (X, C_j, C_k) . We use 0.35 as the threshold to split S_{div} , with $S_{\text{div}1}$ containing all the tuples satisfying $d(j, k) \leq 0.35$, and $S_{\text{div}2}$ containing other tuples. The Pearson correlation of ΔD and Δh on the two subsets are listed in Table 7. According to the results, the correlation is high on $S_{\text{div}1}$ but almost zero on $S_{\text{div}2}$. This is consistent with our intuition that candidates with larger Div scores tend to have higher quality when the distances are under a threshold. However, increasing Div scores does not improve quality when the distances exceed a threshold.

	Twitter-Para		BQ-Para	
	$S_{\text{div}1}$	$S_{\text{div}2}$	$S_{\text{div}1}$	$S_{\text{div}2}$
#num	192	3876	290	6217
ρ	0.635	0.021	0.655	0.025

Table 7: Pearson correlation of ΔD and Δh on two partitions of S_{div} controlled by a threshold (0.35).

The correlation between ΔM and Δh on $S_{\text{div}1}$ are shown in Table 8. It is shown that the correlation scores for all the metrics (except for the $Dist$ function itself) are negative, which means the metrics tend to have opposite judgments with human annotators about the paraphrasing quality for the candidates in $S_{\text{div}1}$.