Not All Metrics Are Guilty: Improving NLG Evaluation by Diversifying References

Tianyi Tang¹,⁶ *, Hongyuan Lu⁴, Yuchen Eleanor Jiang⁵, Haoyang Huang², Dongdong Zhang², Wayne Xin Zhao¹,⁶**, Tom Kocmi³, Furu Wei²
¹ Gaoling School of Artificial Intelligence, Renmin University of China ² Microsoft Research Asia, China ³ Microsoft ⁴ The Chinese University of Hong Kong ⁵ AIWaves Inc. ⁶ Beijing Key Laboratory of Big Data Management and Analysis Methods

{steventianytang,hongyuanlu}@outlook.com  eleanor.jiang@aiwaves.cn {haohua,dozhang,tomkocmi,fuwei}@microsoft.com  batmanfly@gmail.com

Abstract

Most research about natural language generation (NLG) relies on evaluation benchmarks with limited references for a sample, which may result in poor correlations with human judgements. The underlying reason is that one semantic meaning can actually be expressed in different forms, and the evaluation with a single or few references may not accurately reflect the quality of the model’s hypotheses. To address this issue, this paper presents a simple and effective method, named Div-Ref, to enhance existing evaluation benchmarks by enriching the number of references. We leverage large language models (LLMs) to diversify the expression of a single reference into multiple high-quality ones to cover the semantic space of the reference sentence as much as possible. We conduct comprehensive experiments to empirically demonstrate that diversifying the expression of reference can significantly enhance the correlation between automatic evaluation and human evaluation. This idea is compatible with recent LLM-based evaluation which can similarly derive advantages from incorporating multiple references. We strongly encourage future generation benchmarks to include more references, even if they are generated by LLMs, which is once for all. We release all the code and data at https://github.com/RUCAIBox/Div-Ref to facilitate research.

1 Introduction

Evaluation plays a pivotal role in advancing the research on natural language generation (NLG) (Celikyilmaz et al., 2020; Li et al., 2022). It aims to measure the quality of the generated hypotheses in NLG tasks (e.g., machine translation, text summarization, and image caption) from multiple perspectives, such as accuracy, fluency, informativeness, and semantic consistency. There exist two typical approaches for NLG evaluation, namely human evaluation and automatic evaluation. Human evaluation relies on qualified annotators for a reliable assessment of the generation results of NLG models (Sai et al., 2022). However, it is very costly and time-consuming to conduct large-scale human evaluations, especially for complicated tasks.

To reduce the human cost, researchers have proposed various automatic evaluation metrics. Yet, due to their rigid analytic forms, they often suffer from an inaccurate approximation of the task goal, even having significant discrepancies with human evaluation (Zhang et al., 2023). Despite the widespread concerns about evaluation metrics (Sulem et al., 2018; Alva-Manchego et al., 2021), another seldom discussed yet important factor is the number of reference texts in the evaluation benchmarks. There always exist diverse hypotheses that would satisfy the goal of an NLG task, however, the number of ground-truth references provided by human annotators is often limited in scale. For example, there is only one English ground-truth reference written for a Chinese input sentence in multiple cases, which may result in poor correlations with human judgements.

Table 1: The motivation illustration of our proposed Div-Ref method. For the Chinese-to-English translation, the evaluation scores of BLEU and BERTScore are relatively low when using the single ground-truth reference. After diversify the ground truth into multiple references, the correlation of these two metrics with human evaluation can be improved.

<table>
<thead>
<tr>
<th>Input x</th>
<th>精果是我最喜欢的水果。但香蕉是她的最爱。</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference y*</td>
<td>The apple is my most loved fruit but the banana is her most loved.</td>
</tr>
<tr>
<td>Hypothesis y₁</td>
<td>My favorite fruit is apple, while hers beloved is banana.</td>
</tr>
<tr>
<td>BLEU(y₁</td>
<td>y*) = 0.014,  BERTScore(y₁</td>
</tr>
<tr>
<td>Diversified Hypothesis y₁, y₂, y₃</td>
<td>Apple is my favorite fruit, but bananas hold that title for her. My most loved fruit is the apple, while her most beloved is the banana.</td>
</tr>
<tr>
<td>BLEU(y₁</td>
<td>y*, y₂</td>
</tr>
</tbody>
</table>

* This work was done during internship at MSRA.
** Corresponding author

6596
June 16-21, 2024 ©2024 Association for Computational Linguistics
2 Related Work

2.1 Automatic Evaluation

Automatic evaluation metrics for natural language generation could be mainly categorized into two streams: reference-based and reference-free evaluation. The former involves measuring the quality of the hypothesis by comparing it with single or few ground-truth references, e.g., BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005). They primarily focus on the n-gram overlaps between the hypothesis and the references. Recently, neural metrics have become a mainstream method to evaluate semantic similarity and usually have a higher correlation with human evaluation. The representative metrics include BERTScore (Zhang et al., 2020), BLEURT (Sellam et al., 2020), and recent methods involving LLMs (Kocmi and Federmann, 2023; Wang et al., 2023; Chiang and Lee, 2023; Luo et al., 2023; Lu et al., 2023; Gao et al., 2023). Reference-free evaluations assess the hypothesis without the necessity of any reference. They often adopt neural-based models as a black box for evaluating semantic quality as well as grammatical fluency (Zhao et al., 2020; Mehri and Eskenazi, 2020; Hessel et al., 2021; Liu et al., 2023; Chen et al., 2023). However, the reference-free metrics has lower correlation with human compared to the reference-based ones (Kocmi and Federmann, 2023; Wang et al., 2023). In this work, we primarily focus on the reference-based automatic metrics, even without the need for altering their core implementation.

2.2 Increasing the Reference Number

Initially, researchers attempt to utilize paraphrasing methods (Bandel et al., 2022) to enrich the instances of training set (Zheng et al., 2018; Khayrallah et al., 2020). Zhou et al. (2006b) use paraphrasing to enhance the evaluation of the summarization task. There are also prior works that employed paraphrasing in enhancing evaluations with machine translation, either by human paraphrasing (Gupta et al., 2019; Freitag et al., 2020b,a) or automatic paraphrasing (Zhou et al., 2006a; Kauchak and Barzilay, 2006; Thompson and Post, 2020a; Bawden et al., 2020b,a). One recent study reports that the maximization of diversity should be favored for paraphrasing (Bawden et al., 2020b), which enhances the succeeding evaluation. Although current work showcases the promise of paraphrasing methods, they are confined to improving the correlation of specific metrics (e.g., BLEU and ROUGE) in certain tasks (e.g., translation and summarization). They neglect to explore the importance of the number of references, considering constraints such as the quality of automatic paraphrasing or the expense of human paraphrasing. Meanwhile, our investigation reveals that the majority of newly proposed NLG benchmarks in 2023 continue to rely on only one reference. Even those benchmarks...
incorporating multiple references typically feature no more than two or three ground truth. The advent of LLMs has facilitated a convenient and effective means of diversifying references to encompass the semantic space of samples. In this work, we design dedicated prompts tailored for LLMs and extensively investigate the imperative of augmenting the number of references in NLG benchmarks.

3 Methodology

This section first provides a formal definition by introducing several crucial aspects of NLG evaluation. We then describe our approach that leverages LLMs to enrich the semantic coverage of references, bridging the gap between automatic evaluation and human evaluation.

3.1 NLG Evaluation Formulation

As for an NLG task, let \( x \) denote the input sequence associated with extra information (task goal, additional context, etc) and \( y^* \) denote the ground-truth reference provided by the benchmark. After a model or system generates the hypothesis sequence \( \hat{y} \), the automatic evaluation of the metric \( M \) can be represented as \( M(\hat{y}|x, y^*) \). Accordingly, we can also represent human evaluation as \( H(\hat{y}|x, y^*) \).

Hence, to access the quality of the metric \( M \), researchers usually calculate the correlation score with human evaluation \( H \):

\[
\rho(M(\hat{y}|x, y^*), H(\hat{y}|x, y^*)),
\]

where \( \rho \) can be any correlation function such as Spearman correlation and Kendall’s tau. An ideal metric is to maximize the correlation between automatic evaluation \( M \) and human evaluation \( H \).

Note that, \( H \) is a subjective process and cannot be directly calculated. Intuitively, when a human assesses on the hypothesis \( \hat{y} \), he or she will match \( \hat{y} \) among various valid sentences, which can be illustrated as a semantic sentence space \( \mathcal{Y} \) formed in our brain based on human knowledge and common sense related to the ground-truth reference \( y^* \). Therefore, the human evaluation can be further described as \( H(\hat{y}|x, \mathcal{Y}) \).

While researchers on NLG evaluation focus on proposing various implementations of \( M \), we aim to improve the automatic evaluation benchmark using \( M(\hat{y}|x, A(\mathcal{Y})) \), where \( A(\mathcal{Y}) \) is the approximation of \( \mathcal{Y} \) to instantiate the semantic space. \( A(\mathcal{Y}) \) is defined as \( \{y^*, \tilde{y}_1, \ldots, \tilde{y}_n\} \) to alleviate the bias and insufficiency of a single reference in representing the entire semantic space of the ground-truth references. To achieve this, we augment the reference with diverse expressions while retaining the same meaning, aiming to approximate the semantic space \( \mathcal{Y} \). In the traditional single-reference evaluation benchmark, \( A(\mathcal{Y}) \) corresponds to \( \{y^*\} \).

As the acquisition of \( A(\mathcal{Y}) \) is costly for human annotation, we propose to leverage the superior capability of LLMs to generate high-quality and diverse references. With this approach, the automatic evaluation can be formulated as follows:

\[
M(\hat{y}|x, A(\mathcal{Y})) = M(\hat{y}|x, y^*, \tilde{y}_1, \ldots, \tilde{y}_n).
\]

(2)

Traditional metrics, such as BLEU (Papineni et al., 2002) and ChrF (Popović, 2015), have built-in algorithms to handle multiple references, while for neural metrics, they only support a single reference and then aggregate the scores from each reference. In practice, the evaluation score under the multiple-reference setting can be calculated as follows:

\[
M(\hat{y}|x, y^*, \tilde{y}_1, \ldots, \tilde{y}_n) = \mathcal{F} \left[ M(\hat{y}|x, \tilde{y}_i) \right],
\]

(3)

where \( \tilde{y}_0 = y^* \) and \( \mathcal{F} \) is a function leveraged to aggregate scores of multiple diversified sentences, which can be the operation of max aggregation or mean aggregation.

3.2 LLM Diversifying for Evaluation

Recently, LLMs have showcased remarkable capabilities across various NLP tasks. They have proven to be powerful aids in tasks such as text paraphrasing, text style transfer, and grammatical error correction (Kaneko and Okazaki, 2023). Therefore, we harness the potential of LLMs as the approximation function \( A \) to generate diverse expressions \( \tilde{y}_1, \ldots, \tilde{y}_n \) while preserving the original semantics of the ground-truth reference \( y^* \).

3.2.1 Paraphrasing Prompt

Following existing work (Bawden et al., 2020b), we provide the LLM with the paraphrasing prompt “Paraphrase the sentences: \( \{reference\} \)” to wrap the given reference and employ nucleus sampling (Holtzman et al., 2020) to generate a variety of rephrased sentences. In our preliminary experiments, we apply the paraphrasing prompt to paraphrase ten sentences for each English reference sentence from the WMT22 Metrics Shared Task (Freitag et al., 2022). We calculate a semantic diversity score \(^1\) of the rephrased sentences as 0.032.

---

\(^1\)We calculate the mean cosine distance between each rephrased pair using OpenAI Embeddings.
We further observe that rephrased sentences primarily involve word-level substitutions, with minimal modifications to the sentence structure.

### 3.2.2 Diversified Prompts

To improve the diversity of the reference sentences as suggested by Bawden et al. (2020b), we explore several heuristic rules to obtain more diverse texts and cover the semantic space. Inspired by Jiao et al. (2023), we ask ChatGPT to provide instructions that cover different aspects of semantic expressions with the prompt: “Provide ten prompts that can make you diversify the expression of given texts by considering different aspects.”. According to the suggestions by Savage and Mayer (2006), we screen out ten diversifying instructions to promote the changes in words, order, structure, voice, style, etc., which are listed as follows:

1. Change the order of the sentences:
2. Change the structure of the sentences:
3. Change the voice of the sentences:
4. Change the tense of the sentences:
5. Alter the tone of the sentences:
6. Alter the style of the sentences:
7. Rephrase the sentences while retaining the original meaning:
8. Use synonyms or related words to express the sentences with the same meaning:
9. Use more formal language to change the level of formality of the sentences:
10. Use less formal language to change the level of formality of the sentences:

Then, we also utilize the ten instructions to generate ten diversified sentences in total (i.e., one for each instruction). The semantic diversity score increases from 0.032 to 0.049, which demonstrates a significant diversity improvement among the sentences and verifies the effectiveness of our diverse prompts. Note that, LLM diversifying is simple and convenient and does not need any post manual filtering. We conduct further experiments to verify it in Section 4.3.

### 4 Experiments

In this section, we deliberately select three different types of natural language generation tasks to verify the effectiveness of multiple references.

#### 4.1 Experimental Setup

##### 4.1.1 Benchmarks

We choose three meta evaluation benchmarks covering multilingual and multimodal scenarios. These metric benchmarks consist of human scores of the generated text (i.e., $H(y'|x, Y)$), and we can calculate their correlation with the automatic metric scores $M(y'|x, A(\mathcal{Y}))$ using multiple references.

- **WMT22 Metrics Shared Task** (Freitag et al., 2022) includes the generated sentences of different competitor models in the WMT22 News Translation Task (Kocmi et al., 2022). They require human experts to rate these sentences via the multidimensional quality metrics (MQM) schema. We use all three evaluated language pairs, including Chinese (Zh)$\rightarrow$English (En), English (En)$\rightarrow$German (De), and English (En)$\rightarrow$Russian (Ru). We leverage the standardized toolkit mt-metrics-eval V2\(^2\) to calculate the segment-level Kendall Tau score and the

\(^2\)github.com/google-research/mt-metrics-eval
system-level pairwise accuracy following Kocmi et al. (2021). Note that the overall system-level pairwise accuracy across three languages is the most important metric for translation evaluation (Deutsch et al., 2023).

- SummEval (Fabbri et al., 2021) comprises 200 summaries generated by each of the 16 models on the CNN/Daily Mail dataset (See et al., 2017). Human judgements measure these summaries in terms of coherence, consistency, fluency, and relevance. We apply the sample-level Spearman score to measure the correlation.

- PASCAL-50S (Vedantam et al., 2015) is a triple collection of 4,000 instances wherein each instance consists of one reference and two captions. Human annotators compare the two captions based on the reference and express their preference. We calculate the accuracy of whether the metric assigns a higher score to the caption preferred by humans. Our experiments follow the setups outlined by Hessel et al. (2021).

### 4.1.2 Metrics

We evaluate a variety of automatic metrics covering different categories. Based on the taxonomy of existing work (Sai et al., 2022), we select 17 metrics subdivided into five classes:

- Character-based metrics: ChrF (Popović, 2015);
- Word-based metrics: BLEU (Papineni et al., 2002), ROUGE-1/2/L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016);
- Embedding-based metrics: BERTScore (Zhang et al., 2020) and MoverScore;
- Trained metrics: BLEURT (Sellam et al., 2020), Prism (Thompson and Post, 2020b), COMET (Rei et al., 2020), and BARTScore (Yuan et al., 2021);
- LLM-based metrics: GEMBA-Dav3-DA (Kocmi and Federmann, 2023) and ChatGPT-eval (Stars w/ ref) (Wang et al., 2023);

The implementation of each metrics are detailed Appendix A.1. The metrics we used for each benchmark are listed in Table 2.

### 4.1.3 Implementation Details

As for our approach, we utilize the gpt-3.5-turbo-instruct model as the LLM along with the instructions outlined in Section 3.2 to diversify the reference sentences into different expressions. When utilizing the OpenAI API, we set the temperature to 1 and the top_p to 0.9. In Equation 3, we employ the max aggregation and generate 10 diversified sentences (i.e., one for each instruction). We further analyze these hyper-parameters in Section 4.3.

In our experiments, the baseline method is the evaluation of various metrics over single-reference benchmarks, represented by Single-Ref, and the evaluation of our approach over multiple diversified references is denoted as Div-Ref.

### 4.2 Experimental Results

The results of the three evaluation benchmarks over various automatic metrics are shown in the following subsections. We can see that enriching the number of references using our our LLM diversifying method shows a better correlation with human evaluation than the single-reference baseline. Our method is also compatible with existing SOTA LLM-based methods and can enhance them to achieve a higher correlation.

#### 4.2.1 Evaluation on Machine Translation

As shown in the figure 1, our Div-Ref method has shown consistent correlation improvements across all evaluation on the system-level accuracy when compared to the single-reference of the baseline system. Surprisingly, the SOTA metric GEMBA can still be enhanced when evaluated with more references. In terms of different languages, we observe that the diversifying methods are effective

<table>
<thead>
<tr>
<th>Categories</th>
<th>Metrics</th>
<th>Translation</th>
<th>Summarization</th>
<th>Caption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character</td>
<td>ChrF</td>
<td>✓</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>BLEU</td>
<td>✓</td>
<td>–</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>ROUGE-1</td>
<td>–</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>ROUGE-2</td>
<td>–</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>ROUGE-L</td>
<td>–</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>METEOR</td>
<td>–</td>
<td>–</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>CIDEr</td>
<td>–</td>
<td>–</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>SPICE</td>
<td>–</td>
<td>–</td>
<td>✓</td>
</tr>
<tr>
<td>Word</td>
<td>BERTScore</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>MoverScore</td>
<td>–</td>
<td>✓</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>BLEURT</td>
<td>✓</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Prism</td>
<td>✓</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>COMET</td>
<td>✓</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>BARTScore</td>
<td>✓</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>LLM</td>
<td>GEMBA</td>
<td>✓</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>ChatGPT-eval</td>
<td>–</td>
<td>✓</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 2: The summary of metrics evaluated on tasks.
across different languages. English and Russian references benefit more than the German ones, which may be due to the distinct multilingual ability of gpt-3.5-turbo. Notably, our approach showcases significant effects on the traditional BLEU metric, which can further facilitate the application due to its efficiency and universality. The large improvement further demonstrates the automatic metric may be not guilty but the evaluation benchmark needs more references.

4.2.2 Evaluation on Text Summarization
In the summarization task, we select six metrics to examine the correlation against human evaluation
from four aspects: coherence, consistency, fluency, and relevance. According to the results shown in Figure 2, the Div-Ref method can make significant improvements in almost all dimensions compared to the traditional single-reference approach. We can see that the traditional word-based metrics (e.g., ROUGE) and the embedding-based metrics (e.g., BERTScore) perform closely, while LLM-based metric shows remarkable correlation with human evaluation. This phenomena further demonstrates the effectiveness of LLMs for NLG evaluation, as described by Wang et al. (2023). It should be noted that our method has further improved the LLM-based metric ChatGPT-eval in all dimensions. This also shows that our approach is effective in improving the correlation with human evaluation and the NLG benchmarks should include more references.

4.2.3 Evaluation on Image Caption

In order to examine the effectiveness of our method for the image caption task, we expand the reference under four different settings to judge whether the metric assigns a higher score to the caption preferred by humans. The results of the image caption task are reported in Figure 3. For the HC and MM settings, which are difficult settings to judge two similar captions, Div-Ref exhibits enhancements in all metrics, particularly for SPICE, METEOR, and BERTScore. This verifies our approach can expand the semantic coverage of references to bridge the gap between automatic evaluation and human evaluation. Regarding HI and HM, Div-Ref still maintains the improvements in all metrics, except for a slight drop for BERTScore in the HM setting. Despite one of the candidate captions being incorrect or machine-generated, our method can strongly align different metrics with human preference, particularly for the SPICE metric. In comparison to the single-reference baseline, our approach yields a significant improvement of 3.6 points with SPICE in HI and 2.9 points for HM.

4.3 Ablation Analysis

In this section, we examine the impact of various factors of increasing the reference numbers, which include the selection of diversifying models, the application of instruction prompts, the choice of the aggregation function, the effect of post-filtering, and the number of diversified references. The results can be found in Table 3 and 4 and Figure 4.

(1) Firstly, we compare the influence of our diversifying LLM gpt-3.5-turbo-instruct with three rephrasing PLMs PEGASUS-Paraphrasing3.

³https://huggingface.co/tuner007/pegasus_paraphrase
Table 3: Analysis of the effect of the diversifying models, instruction prompts, aggregation functions, and post-filtering. We report the system-level accuracy and segment-level correlation of the Chinese-to-English direction over the WMT22 Metric Task. × of PEGASUS, Parrot, and QCPG denotes the three methods do not support multilingual scenario. × of “Built-in” means the metric do not have built-in multi-reference aggregation option. – in “Multilingual” represents the multilingual diverse prompt has the same results as the English diverse prompt.

Table 4: Ablation analysis in the English-to-German and English-to-Russia and directions using segment-level Kendall Tau correlation.

Parrot⁴, and QCPG (Bandel et al., 2022), which are fine-tuned on paraphrasing tasks. However, these three models only support English paraphrasing. We also incorporate another open-source LLMs, LLaMA-2-70b-chat, to diversify our references. From the results, we observe that gpt-3.5-turbo-instruct can outperform three PLMs and LLaMA-2-chat in all metrics, which showcases its superior capability in completing the semantic space of given reference.

(2) Regarding the choice of instruction prompts, we first degrade the diverse prompts to the basic prompt mentioned in Section 3.2. We observe that the diverse prompts can achieve satisfactory results on English references (i.e., Zh-En), and may slightly reduce the performance on non-English languages (Table 4). Then, we further translate the English diverse prompts into respective language (i.e., instructing LLMs using the reference language), and find the gains of multilingual diverse prompts are also not obvious. We attribute the two results to that fact the diversifying ability of LLMs in non-English is not as good as that in English, since English is the dominant language. Besides, we analyze each kind of our diverse prompts in Appendix. We discover that when changing the aggregation and the built-in multi-reference aggregation of BLEU and ChrF, their performances are indistinguishable.

(3) Thirdly, we investigate the aggregation functions using the mean aggregation and the built-in multi-reference aggregation of BLEU and ChrF. We discover that when changing the aggregation from max to mean, the correlation scores for most metrics have dropped, especially in the Chinese-to-English direction. This indicates that the highest-quality reference plays a dominant role in generation evaluation, and our approach to increasing the number of references significantly strengthens this probability. However, averaging multiple reference scores could introduce noise from low-quality reference scores. As for the built-in method of BLEU and ChrF, their performances are indistinguishable.

⁴https://huggingface.co/prithivida/parrot_paraphraser_on_T5
Table 5: Diverse prompts analysis in the Chinese-to-English direction using segment-level Kendall Tau correlation.

In addition, we attempt to filter the generated references considering some of them may be of low quality. We employ gpt-3.5-turbo to judge using the instruction: “Sentence 1: {ref}
Sentence 2: {div_ref}Do sentence 1 and sentence 2 convey the same meaning?”. After eliminating the reference unrecognized by gpt-3.5-turbo, we can find that the removal of low-quality sentences has minimal impact on correlation results. We speculate that our approach involves aggregating results from multiple references and selecting the one with the highest score, effectively disregarding those of inferior quality.

Finally, we examine the influence of scaling the number of references. We utilize the diverse prompts to generate more references. From Figure 4, we observe a consistent upward trend in the overall performance as the number of references increases. For word-based metrics, this growth trend is more obvious. This experiment further shows that traditional benchmarks that rely on a single reference is very one-sided for NLG evaluation, and we need to provide multiple references for benchmarks. Considering that the performance of neural metrics tends to saturate when the quantity is high, over-generation may not lead to more significant gains, suggesting that the optimal cost-effective number may not exceed 20.

5 Conclusion

In this paper, we have investigated the effect of enriching the number of references in NLG benchmarks and verified its effectiveness. Our diversifying method, Div-Ref, can effectively cover the semantic space of the golden reference, which can largely extend the limited references in existing benchmarks. With extensive experiments, our approach yields substantial improvements in the consistencies between evaluation metrics and human evaluation. In future work, we will explore the current evaluation method on more NLG tasks, and also consider extending it to evaluate generation tasks in other modalities. It is also valuable to investigate whether paraphrasing can improve LLMs’ training and utilization.

Acknowledgement

This work was partially supported by Beijing Natural Science Foundation under Grant No. L233008 and 4222027. Xin Zhao is the corresponding author.
Limitations

Despite conducting numerous experiments, further research is required to explore the number of references and the optimal diversifying techniques that can achieve a trade-off between time and effectiveness. Since using more references leads to more evaluation time, future work can explore strategies for mitigating these issues, possibly through the implementation of a selection mechanism that prioritizes sentences with diverse expressions while minimizing the overall number of reference sentences. Moreover, our diverse prompts may fail in specialized domains, such as finance and biomedicine. Rewriting professional terms may lead to inaccuracy evaluation of the generated sentences. Future work can further investigate and validate the effectiveness of our method within these domains. Additionally, we can design more fine-grained prompts tailored to address the specific challenges posed by professional terminology. In addition, due to the high cost of text-davinci-003, we omit the experiments of GEMBA in the ablation analysis, which may lead to an incomplete analysis of LLM-based metrics. The OpenAI API also is non-deterministic, which may lead to different diversifying results for the same input. There is also a chance that OpenAI will remove existing models.

References


A Experimental Details

A.1 Metric Implementation

The implementation details of each metric in different benchmarks are listed as follows:

- **ChrF (Popović, 2015):** We utilize sentence-level ChrF from SacreBLEU for machine translation.
- **BLEU (Papineni et al., 2002):** We utilize sentence-level BLEU from SacreBLEU for machine translation, and employ BLEU from pycocoevalcap for image caption.
- **ROUGE-1/2/L (Lin, 2004):** We utilize ROUGE-1/2/L from files2rouge for text summarization, and employ ROUGE-L from pycocoevalcap for image caption.
- **METEOR (Banerjee and Lavie, 2005):** We utilize METEOR from pycocoevalcap for image caption.
- **CIDEr (Banerjee and Lavie, 2005):** We utilize CIDEr from pycocoevalcap for image caption.
- **SPICE (Banerjee and Lavie, 2005):** We utilize SPICE from pycocoevalcap for image caption.
- **BERTScore (Zhang et al., 2020):** We utilize BERTScore from its official repository (i.e., German and Russia).
- **MoverScore (Zhao et al., 2019):** We utilize MoverScore from its official repository for text summarization. Specially, we leverage the MNLI-BERT checkpoint.
- **BLEURT (Sellam et al., 2020):** We utilize BLEURT from its official repository for machine translation. Specially, we leverage the BLEURT-20 checkpoint.
- **Prism (Thompson and Post, 2020b):** We utilize Prism from its official repository for machine translation.
- **COMET (Rei et al., 2020):** We utilize COMET from its official repository for machine translation. Specially, we leverage the Unbabel/wmt22-comet-da checkpoint.
- **BARTScore (Yuan et al., 2021):** We utilize BARTScore from its official repository for machine translation in the Chinese-to-English direction. Specially, we leverage the BARTScore+CNN+Para checkpoint.
- **GEMBA (Kocmi and Federmann, 2023):** We utilize GEMBA-Dav3-DA from its official repository for machine translation. Specially, we leverage direct assessment as the scoring task, and apply text-davinci-003 as the evaluation model with temperature=0.
- **ChatGPT- eval (Wang et al., 2023):** We utilize ChatGPT- eval (Stars w/ ref) from its official repository for text summarization. Specially, we leverage the star prompt with reference, and apply gpt-3.5-turbo as the evaluation model with temperature=0.

A.2 Diversified Examples

---

5https://github.com/mjpost/sacrebleu
6https://github.com/mjpost/sacrebleu
7https://github.com/salaniz/pycocoevalcap
8https://github.com/pltrdy/files2rouge
9https://github.com/salaniz/pycocoevalcap
10https://github.com/Tiiiger/bert_score
11https://github.com/AIPHES/emnlp19-moverscore
12https://github.com/google-research/bleurt
13https://github.com/thompsonb/prism
14https://github.com/Unbabel/COMET
15https://github.com/neulab/BARTScore
16https://github.com/MicrosoftTranslator/GEMBA
17https://github.com/krystalan/chatgpt_as_nlg_evaluator
### Table 6: The diversified example of WMT22 Metrics Task in the Chinese-to-English direction. More examples can be found at [https://github.com/RUCAIBox/Div-Ref](https://github.com/RUCAIBox/Div-Ref).

<table>
<thead>
<tr>
<th>Source</th>
<th>是否有途径处罚他</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ground-truth reference</strong></td>
<td>Is there a way to punish him?</td>
</tr>
<tr>
<td><strong>Diversified references</strong></td>
<td>Can he be penalized?</td>
</tr>
</tbody>
</table>

### Table 7: The diversified example of WMT22 Metrics Task in the English-to-German direction. More examples can be found at [https://github.com/RUCAIBox/Div-Ref](https://github.com/RUCAIBox/Div-Ref).

<table>
<thead>
<tr>
<th>Source</th>
<th>Ich hoffe wirklich, dass Sie eine Lösung finden werden</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ground-truth reference</strong></td>
<td>Ich hoffe wirklich, dass Sie eine Lösung finden werden</td>
</tr>
</tbody>
</table>
Source
I see it all the time in my line of work.

Ground-truth reference
Я постоянно вижу такое в своей сфере деятельности.

Diversified references
Я всегда наблюдаю подобное в своей сфере работы.  
Такое я вижу постоянно в своей сфере деятельности.  
Такое я постоянно вижу в своей сфере деятельности.  
Такое постоянно видится мной в моей сфере деятельности.  
Я постоянно увижу такое в своей сфере деятельности.  
В своей сфере деятельности я часто наблюдаю подобное.  
Я всегда наблюдаю подобное в своей сфере работы.  
В своей сфере деятельности я непрерывно наблюдаю подобное.  
Я постоянно наблюдаю подобные вещи в своей сфере профессиональной деятельности.  
Я всегда это наблюдаю в своей работе.

Table 8: The diversified example of WMT22 Metrics Task in the English-to-Russian direction. More examples can be found at https://github.com/RUCAIBox/Div-Ref.