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

Tianyi Tang[1,6 *], **Hongyuan Lu**[4], **Yuchen Eleanor Jiang**[5], **Haoyang Huang**[2], **Dongdong Zhang**[2], **Wayne Xin Zhao**[1,6 ✉], **Tom Kocmi**[3], **Furu Wei**[2]

[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
`{steventianyitang,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 aspects, 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.

Input $\mathbf{x}$	苹果是我最喜欢的水果，但香蕉是她的最爱。
Reference $\mathbf{y}^*$	The apple is my most loved fruit but the banana is her most loved.
Hypothesis $\hat{\mathbf{y}}$	My favorite fruit is apple, while hers beloved is banana.
BLEU$(\hat{\mathbf{y}}\vert\mathbf{y}^*) = 0.014$**,**	**BERTScore**$(\hat{\mathbf{y}}\vert\mathbf{y}^*) = 0.923$
Diversified references $\tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2, \tilde{\mathbf{y}}_3$	Apples rank as my favorite fruit, but bananas hold that title for her. Apple is my favorite fruit, but banana is her most beloved. My most loved fruit is the apple, while her most loved is the banana.
BLEU$(\hat{\mathbf{y}}\vert\mathbf{y}^*, \tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2, \tilde{\mathbf{y}}_3) = 0.251$**,**	**BERTScore**$(\hat{\mathbf{y}}\vert\mathbf{y}^*, \tilde{\mathbf{y}}_1, \tilde{\mathbf{y}}_2, \tilde{\mathbf{y}}_3) = 0.958$

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.

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

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

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6596–6610
June 16-21, 2024

the WMT22 News Translation Task (Kocmi et al., 2022). This potentially leads to unreliable evaluation results when using limited ground-truth references, as illustrated in Table 1.

Considering the above-mentioned issue, this paper attempts to improve the NLG evaluation benchmarks and make existing automatic metrics better reflect the actual quality of the hypotheses. We focus on increasing the number of reference texts to narrow the gap between automatic and human evaluation. The key idea is to leverage the excellent ability of existing LLMs to provide more high-quality references for a single sample. By enriching the diversity of the references while maintaining semantic consistency, we expand the coverage of the semantic expressions for evaluating the generated texts from *a single or few* standard references to *a more diverse set* of semantically equivalent references. In this way, our evaluation method can better approximate human evaluation criteria, as the improved scores shown in Table 1. In addition, increasing the number of references is *agnostic* to specific task settings and can be integrated with various automatic metrics for evaluating different generation tasks.

To demonstrate the effectiveness of diversifying references, we conduct extensive experiments on the benchmarks from multiple NLG tasks. The experimental results demonstrate that incorporating multiple references can significantly improve the consistency between traditional evaluation metrics and human evaluation results. Surprisingly, it is even applicable in multilingual and multimodal text generation scenarios. Importantly, our approach is orthogonal with automatic metrics, enabling even the recent LLM-based evaluations (Kocmi and Federmann, 2023; Wang et al., 2023) to benefit from our diversified references and achieve SOTA correlation with human judges. *Therefore, incorporating more references for the NLG benchmark proves advantageous, requiring a one-time effort, and future researchers can reap its benefits.*

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 $\mathbf{x}$ denote the input sequence associated with extra information (task goal, additional context, *etc*) and $\mathbf{y}^*$ denote the ground-truth reference provided by the benchmark. After a model or system generates the hypothesis sequence $\hat{\mathbf{y}}$, the automatic evaluation of the metric $\mathcal{M}$ can be represented as $\mathcal{M}(\hat{\mathbf{y}}|\mathbf{x}, \mathbf{y}^*)$. Accordingly, we can also represent human evaluation as $\mathcal{H}(\hat{\mathbf{y}}|\mathbf{x}, \mathbf{y}^*)$. Hence, to access the quality of the metric $\mathcal{M}$, researchers usually calculate the correlation score with human evaluation $\mathcal{H}$:

$$\rho(\mathcal{M}(\hat{\mathbf{y}}|\mathbf{x}, \mathbf{y}^*), \mathcal{H}(\hat{\mathbf{y}}|\mathbf{x}, \mathbf{y}^*)), \tag{1}$$

where ρ can be any correlation function such as Spearman correlation and Kendall's tau. An ideal metric is to maximize the correlation between automatic evaluation $\mathcal{M}$ and human evaluation $\mathcal{H}$.

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

While researchers on NLG evaluation focus on proposing various implementations of $\mathcal{M}$, we aim to improve the automatic evaluation benchmark using $\mathcal{M}(\hat{\mathbf{y}}|\mathbf{x}, A(\mathbb{Y}))$, where $A(\mathbb{Y})$ is the approximation of $\mathbb{Y}$ to instantiate the semantic space. $A(\mathbb{Y})$ is defined as $\{\mathbf{y}^*, \tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_n\}$ 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 $\mathbb{Y}$. In the traditional single-reference evaluation benchmark, $A(\mathbb{Y})$ corresponds to $\{\mathbf{y}^*\}$.

As the acquisition of $A(\mathbb{Y})$ 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:

$$\mathcal{M}(\hat{\mathbf{y}}|\mathbf{x}, A(\mathbb{Y})) = \mathcal{M}(\hat{\mathbf{y}}|\mathbf{x}, \mathbf{y}^*, \tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_n). \tag{2}$$

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:

$$\mathcal{M}(\hat{\mathbf{y}}|\mathbf{x}, \mathbf{y}^*, \tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_n) = \mathop{\mathcal{F}}_{i=0}^{n} \left[\mathcal{M}(\hat{\mathbf{y}}|\mathbf{x}, \hat{\mathbf{y}_i})\right], \tag{3}$$

where $\hat{\mathbf{y}_0} = \mathbf{y}^*$ and $\mathcal{F}$ is a function leveraged to aggregate scores of multiple diversified sequences, 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 $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_n$ while preserving the original semantics of the ground-truth reference $\mathbf{y}^*$.

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. d

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, our diversifying method is not just paraphrasing but attempts to cover different aspects of the reference expressions. Considering the strong cross-lingual generation capabilities of LLMs (Muennighoff et al., 2022), we apply English instructions to diversify references in different languages (*e.g.,* German and Russian). The diversified examples can be found in Tables 6, 7, 8.

text-embedding-ada-002. Then, we average the score of each instance to obtain an overall semantic diversity score.

3.2.3 Discussion

Compared with existing work (Freitag et al., 2020b; Bawden et al., 2020b) that utilizes paraphrasing for evaluation, we leverage the recent superior LLMs for diversifying the expressions of given reference. After supervised fine-tuning and reinforcement learning from human feedback, LLMs showcase excellent capability to follow the input instruction and align with human preference, which can not achieve by previous paraphrasing methods. To verify the effectiveness of LLMs, we further conduct experiments in Section 4.3 to compare them with traditional paraphrasing models. Moreover, we conduct experiments to evaluate the diversifying results of LLMs. We employ another excellent GPT 3.5 to judge whether the generated sentence conveys the same meaning of given reference. The results show that 94.6% of the generated sentences are suitable, which demonstrates the effectiveness and robustness 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.,* $\mathcal{H}(\mathbf{y}'|\mathbf{x}, \mathbb{Y})$), and we can calculate their correlation with the automatic metric scores $\mathcal{M}(\mathbf{y}'|\mathbf{x}, A(\mathbb{Y}))$ 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)→English (En), English (En)→German (De), and English (En)→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**.

Categories	Metrics	Translation	Summarization	Caption
Character	ChrF	✓	–	–
Word	BLEU	✓	–	✓
	ROUGE-1	–	✓	–
	ROUGE-2	–	✓	–
	ROUGE-L	–	✓	✓
	METEOR	–	–	✓
	CIDEr	–	–	✓
	SPICE	–	–	✓
Embedding	BERTScore	✓	✓	✓
	MoverScore	–	✓	–
Trained	BLEURT	✓	–	–
	Prism	✓	–	–
	COMET	✓	–	–
	BARTScore	✓	–	–
LLM	GEMBA	✓	–	–
	ChatGPT-eval	–	✓	–

Table 2: The summary of metrics evaluated on tasks.

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

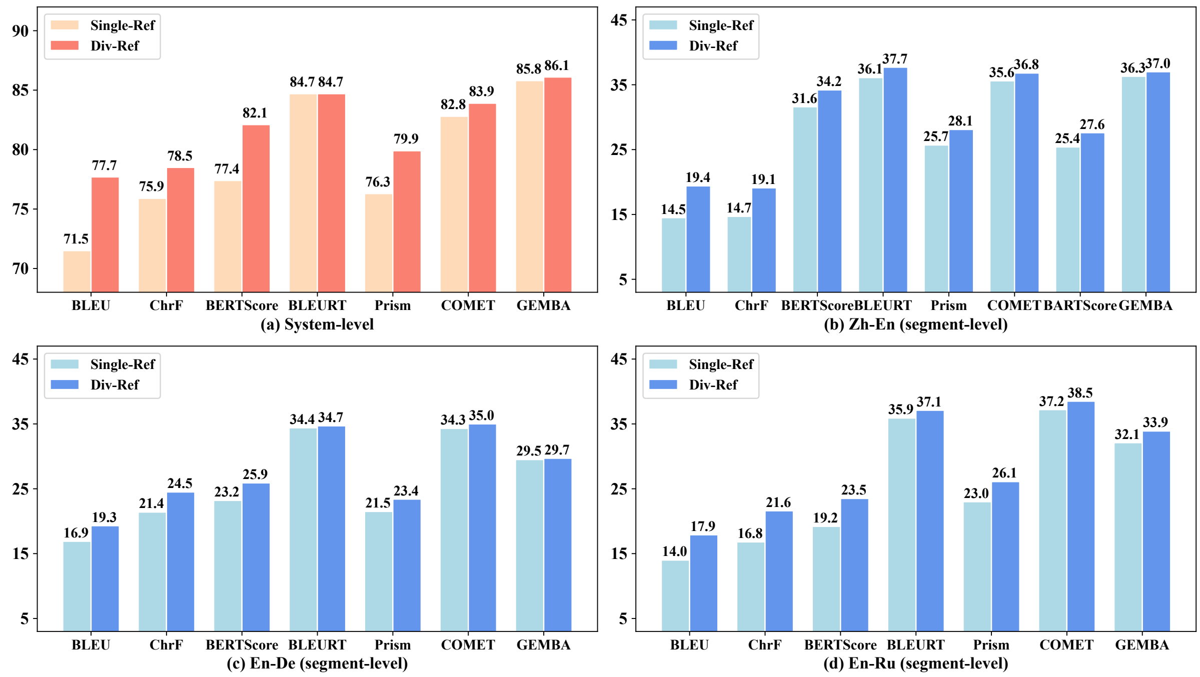

Figure 1: System-level pairwise accuracy (main aspect) and Kendall Tau correlation of segment-level score over the WMT22 Metrics Shared Task on three translation directions.

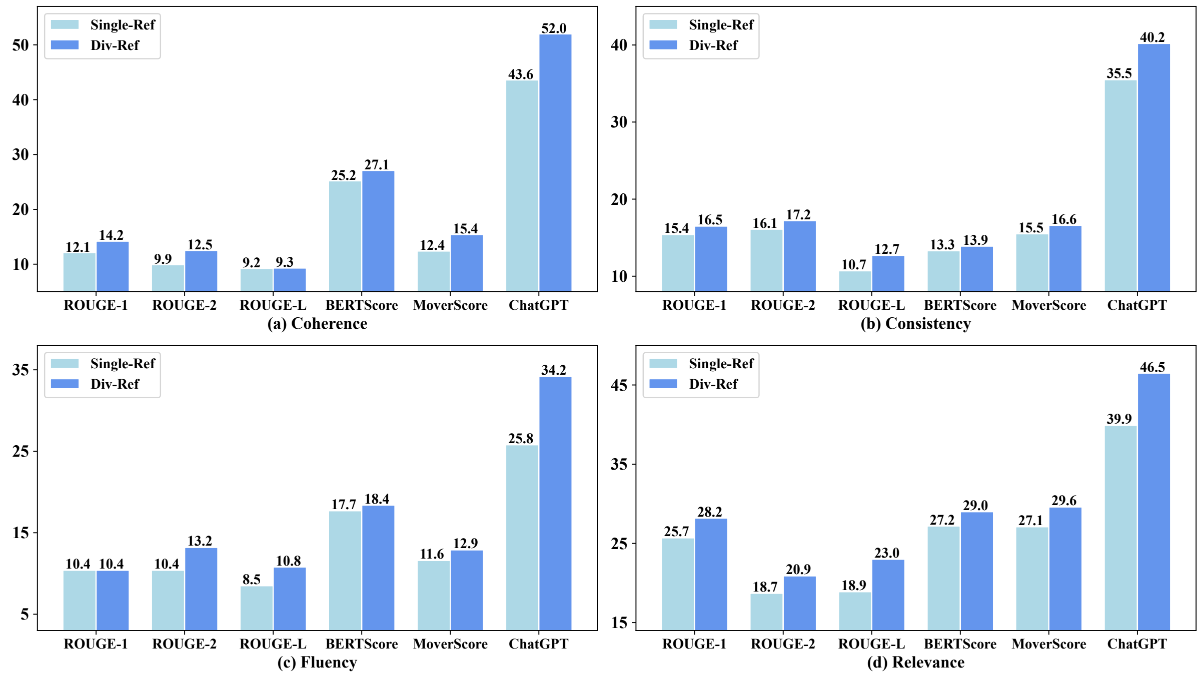

Figure 2: Spearman score of sample-level correlation over the SummEval benchmark on four evaluation aspects.

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

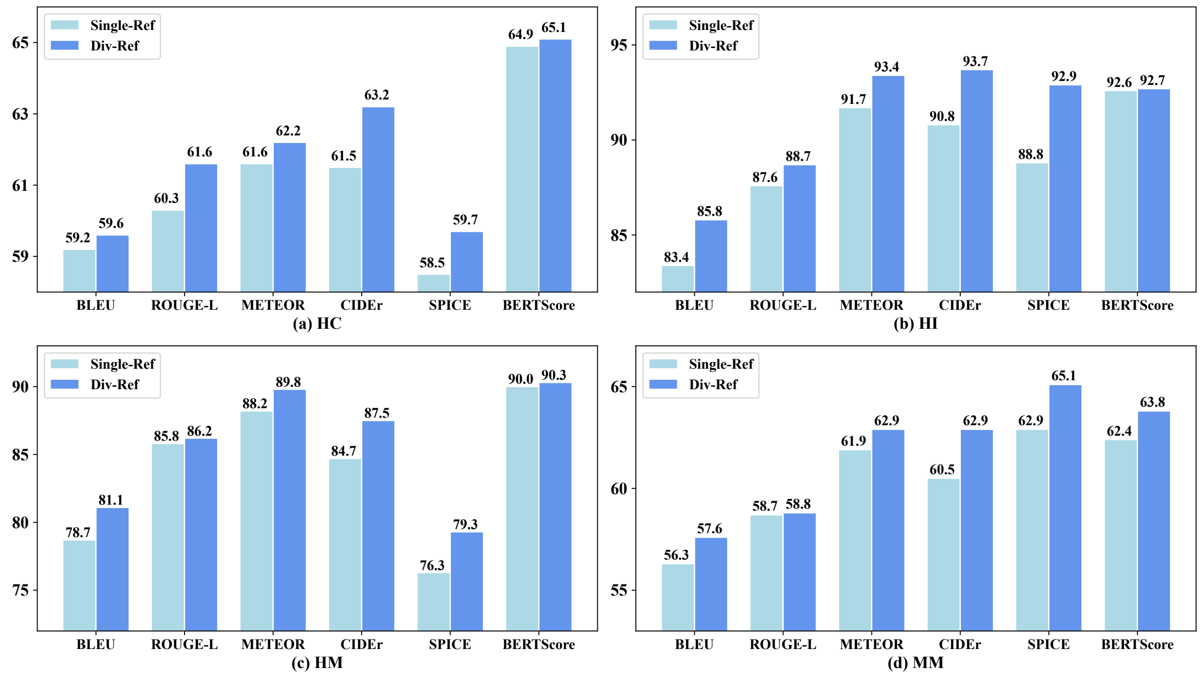

Figure 3: Accuracy score over the PASCAL-50S benchmark on four settings. HC denotes the two captions are correct and written by humans. HI denotes two human-written captions but one is irrelevant. HM denotes one caption is human-written and the other is model-generated. MM denotes two model-generated captions.

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-Paraphrasing[3],

[3] https://huggingface.co/tuner007/pegasus_paraphrase

Settings		BLEU		ChrF		BERTScore		BLEURT		Prism		COMET		Average Gains	
		System	Zh-En	System	Zh-En	System	Zh-En	System	Zh-En	System	Zh-En	System	Zh-En	System	Zh-En
Single-Ref		71.5	14.5	75.9	14.7	77.4	31.6	84.7	36.1	76.3	25.7	82.8	35.6	0.0	0.0
Ours (GPT 3.5+Diverse+Max)		77.7	19.4	78.5	19.1	82.1	34.2	84.7	37.7	79.9	28.1	83.9	36.8	+3.0	+2.9
Model	PEGASUS	×	18.2	×	18.5	×	33.2	×	37.0	×	27.4	×	36.0	×	+2.0
	Parrot	×	17.5	×	18.3	×	32.2	×	36.8	×	26.3	×	36.1	×	+1.5
	QCPG	×	17.4	×	17.2	×	32.8	×	37.0	×	26.8	×	36.2	×	+1.5
	LLaMA-2-70b-chat	74.5	17.5	76.3	16.6	79.2	32.9	83.6	36.8	78.8	26.8	82.5	36.3	+1.1	+1.4
Prompt	Basic	77.4	17.6	77.4	16.9	81.8	33.2	83.9	37.1	79.2	27.1	83.2	36.3	+2.4	+1.7
	Multilingual	77.7	–	77.7	–	81.8	–	84.7	–	79.2	–	83.9	–	+2.7	0.0
Aggregation	Mean	77.0	16.6	78.8	10.5	83.2	32.2	81.8	35.5	79.2	23.1	81.8	33.9	+2.2	-1.1
	Built-in	78.5	18.8	78.5	19.1	×	×	×	×	×	×	×	×	×	×
Filtering subpar references		77.7	19.2	78.5	19.0	82.1	34.1	84.3	37.6	79.9	28.0	83.9	36.8	0.0	-0.1

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 "Bulit-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.

Settings		BLEU		ChrF		BERTScore		BLEURT		Prism		COMET		Average Gains	
		En-De	En-Ru	En-De	En-Ru	En-De	En-Ru	En-De	En-Ru	En-De	En-Ru	En-De	En-Ru	En-De	En-Ru
Single-Ref		16.9	14.0	21.4	16.8	23.2	19.2	34.4	35.9	21.5	23.0	34.3	37.2	0.0	0.0
Ours (GPT 3.5+Diverse+Max)		19.3	17.9	24.5	21.6	25.9	23.5	34.7	37.1	23.4	26.1	35.0	38.5	+1.9	+3.1
Model	LLaMA-2-70b-chat	18.1	16.0	22.8	19.5	24.1	21.6	34.8	36.8	22.4	24.7	35.1	38.2	+0.9	+1.7
Prompt	Basic	19.6	19.3	25.2	24.2	26.2	25.4	35.5	34.7	23.9	23.0	35.2	34.8	+2.3	+2.6
	Multilingual	18.9	19.1	22.4	22.2	23.9	24.2	37.3	37.1	26.4	26.1	38.7	38.9	+2.7	+3.6
Aggregation	Mean	13.9	15.0	17.2	16.3	20.0	19.4	32.3	37.0	19.2	22.3	32.0	36.6	-2.8	+0.1
	Built-in	18.4	18.1	24.5	21.6	×	×	×	×	×	×	×	×	×	×
Filtering subpar references		19.4	17.9	24.8	21.6	26.0	23.5	34.8	37.1	23.4	26.1	35.1	38.5	+0.2	0.0

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

Parrot[4], 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 degrades 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 compare a mixture of one sentence per prompt with ten sentences per prompt. From the results in Table 5, we can find that mixing prompts is better than any individual prompt. This further demonstrates the effectiveness of our delicate prompts and they can cover a broader semantics range of reference sentences.

(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.

[4] https://huggingface.co/prithivida/parrot_paraphraser_on_T5

Prompts	BLEU	ChrF	BERTScore	BLEURT	Prism	COMET	Average Gains
Single-Ref	14.5	14.7	31.6	36.1	25.7	35.6	0.0
Ours (Mixing ①-⑩)	19.4	19.1	34.2	37.7	28.1	36.8	+2.9
① × 10	16.6	16.3	33.0	37.1	26.8	36.3	+1.3
② × 10	15.9	15.5	32.2	36.4	26.4	35.7	+0.6
③ × 10	17.8	17.5	33.0	36.8	27.0	36.2	+1.7
④ × 10	16.8	16.7	32.8	36.9	26.6	36.0	+1.3
⑤ × 10	15.1	15.4	32.0	36.3	26.1	35.6	+0.4
⑥ × 10	18.1	17.5	33.5	37.4	27.4	36.3	+2.0
⑦ × 10	17.4	16.5	33.4	37.2	27.0	36.4	+1.6
⑧ × 10	18.1	17.2	33.4	37.4	27.2	36.4	+1.9
⑨ × 10	16.8	16.2	33.1	37.3	26.8	36.2	+1.4
⑩ × 10	18.6	19.0	33.7	37.2	27.5	36.5	+2.4

Table 5: Diverse prompts analysis in the Chinese-to-English direction using segment-level Kendall Tau correlation.

(4) 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}\nSentence 2: {div_ref}\nDo sentence 1 and sentence 2 convey the same meaning?\n\n*". 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.

(5) 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 relies 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.

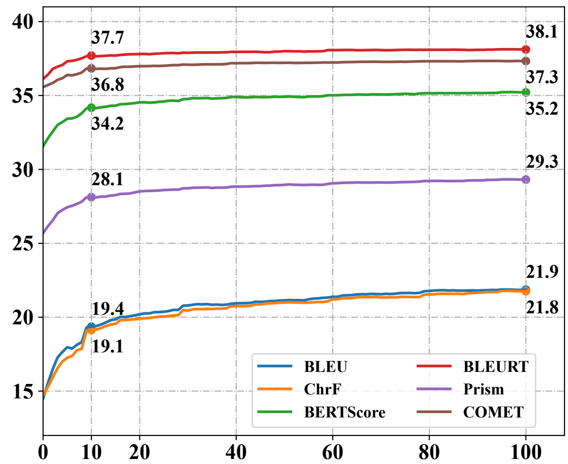

Figure 4: Kendall Tau correlation score *w.r.t.* the number of generated references in the Chinese-to-English direction on the WMT22 Metrics Shared Task.

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

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision – ECCV 2016*, pages 382–398, Cham. Springer International Publishing.

Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim, and Liat Ein-Dor. 2022. Quality controlled paraphrase generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 596–609, Dublin, Ireland. Association for Computational Linguistics.

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Rachel Bawden, Biao Zhang, Andre Tättar, and Matt Post. 2020a. ParBLEU: Augmenting metrics with automatic paraphrases for the WMT'20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 887–894, Online. Association for Computational Linguistics.

Rachel Bawden, Biao Zhang, Lisa Yankovskaya, Andre Tättar, and Matt Post. 2020b. A study in improving BLEU reference coverage with diverse automatic paraphrasing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 918–932, Online. Association for Computational Linguistics.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. *arXiv preprint arXiv:2304.00723*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Modifying kendall's tau for modern metric meta-evaluation. *arXiv preprint arXiv:2305.14324*.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Markus Freitag, George Foster, David Grangier, and Colin Cherry. 2020a. Human-paraphrased references improve neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1183–1192, Online. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020b. 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, Online. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskenazi, and Jeffrey Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 379–391, Stockholm, Sweden. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

WX Jiao, WX Wang, JT Huang, Xing Wang, and ZP Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Masahiro Kaneko and Naoaki Okazaki. 2023. Reducing sequence length by predicting edit operations with large language models. *arXiv preprint arXiv:2305.11862*.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for automatic evaluation. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 455–462, New York City, USA. Association for Computational Linguistics.

Huda Khayrallah, Brian Thompson, Matt Post, and Philipp Koehn. 2020. Simulated multiple reference training improves low-resource machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 82–89, Online. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2022. A survey of pretrained language models based text generation. *arXiv preprint arXiv:2201.05273*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *arXiv preprint arXiv:2303.13809*.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

Shikib Mehri and Maxine Eskenazi. 2020. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, Online. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference* *on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Comput. Surv.*, 55(2).

Alice Savage and Patricia Mayer. 2006. *Effective academic writing: the short essay*. Oxford University Press.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020a. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020b. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, Los Alamitos, CA, USA. IEEE Computer Society.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

Wei Zhao, Goran Glavaš, Maxime Peyrard, Yang Gao, Robert West, and Steffen Eger. 2020. On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1656–1671, Online. Association for Computational Linguistics.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics.

Renjie Zheng, Mingbo Ma, and Liang Huang. 2018. Multi-reference training with pseudo-references for neural translation and text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3188–3197, Brussels, Belgium. Association for Computational Linguistics.

Liang Zhou, Chin-Yew Lin, and Eduard Hovy. 2006a. Re-evaluating machine translation results with paraphrase support. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 77–84, Sydney, Australia. Association for Computational Linguistics.

Liang Zhou, Chin-Yew Lin, Dragos Stefan Munteanu, and Eduard Hovy. 2006b. ParaEval: Using paraphrases to evaluate summaries automatically. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 447–454, New York City, USA. Association for Computational Linguistics.

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[5] for machine translation.
- BLEU (Papineni et al., 2002): We utilize sentence-level BLEU from SacreBLEU[6] for machine translation, and employ BLEU from pycocoevalcap[7] for image caption.
- ROUGE-1/2/L (Lin, 2004): We utilize ROUGE-1/2/L from files2rouge[8] for text summarization, and employ ROUGE-L from pycocoevalcap[9] for image caption.
- METEOR (Banerjee and Lavie, 2005): We utilize METEOR from pycocoevalcap[9] for image caption.
- CIDEr (Banerjee and Lavie, 2005): We utilize CIDEr from pycocoevalcap[9] for image caption.
- SPICE (Banerjee and Lavie, 2005): We utilize SPICE from pycocoevalcap[9] for image caption.
- BERTScore (Zhang et al., 2020): We utilize BERTScore from its official repository[10] for machine translation, text summarization, and image caption. Specially, we leverage roberta-large for English reference sentences, while apply bert-base-multilingual-cased for other languages (*i.e.,* German and Russia).
- MoverScore (Zhao et al., 2019): We utilize MoverScore from its official repository[11] for text summarization. Specially, we leverage the MNLI-BERT checkpoint.
- BLEURT (Sellam et al., 2020): We utilize BLEURT from its official repository[12] for machine translation. Specially, we leverage the BLEURT-20 checkpoint.
- Prism (Thompson and Post, 2020b): We utilize Prism from its official repository[13] for machine translation.
- COMET (Rei et al., 2020): We utilize COMET from its official repository[14] for machine translation. Specially, we leverage the Unbabel/wmt22-comet-da checkpoint.
- BARTScore (Yuan et al., 2021): We utilize BARTScore from its official repository[15] 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[16] 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[17] 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

[5] https://github.com/mjpost/sacrebleu
[6] https://github.com/mjpost/sacrebleu
[7] https://github.com/salaniz/pycocoevalcap
[8] https://github.com/pltrdy/files2rouge
[9] https://github.com/salaniz/pycocoevalcap
[10] https://github.com/Tiiiger/bert_score
[11] https://github.com/AIPHES/emnlp19-moverscore
[12] https://github.com/google-research/bleurt
[13] https://github.com/thompsonb/prism
[14] https://github.com/Unbabel/COMET
[15] https://github.com/neulab/BARTScore
[16] https://github.com/MicrosoftTranslator/GEMBA
[17] https://github.com/krystalan/chatgpt_as_nlg_evaluator

Source
是否有途径处罚他
Ground-truth reference
Is there a way to punish him?
Diversified references
Can he be penalized?
Is there a way to punish him?
Can he be punished in any way?
Can he be punished?
Was there a way to punish him?
Can he be punished in any way?
Is there a means of reprimanding him?
Can he be punished in any way?
Is there a means by which he may be disciplined?
Can we do something to punish him?

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.

Source
I sincerely hope you get to find a resolution
Ground-truth reference
Ich hoffe wirklich, dass Sie eine Lösung finden werden
Diversified references
Ich drücke die Daumen, dass Sie eine Lösung ausarbeiten können.
Ich hoffe wirklich, dass Sie eine Lösung finden werden.
Ich hoffe, dass Sie eine Lösung finden werden, wirklich.
Ich hoffe wirklich, dass eine Lösung von Ihnen gefunden werden wird.
Ich hatte gehofft, dass Sie eine Lösung finden würden.
Ich hoffe, dass Sie eine Lösung finden werden
Ich wünsche Ihnen aufrichtig, dass Sie eine Lösung finden werden.
Ich wünsche mir innigst, dass Sie eine Lösung finden werden.
Ich hoffe aufrichtig, dass Sie eine Lösung finden werden.
Ich hoffe wirklich, dass du eine Lösung findest.

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.

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 за подобным 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.