Beyond Reference: Evaluating High Quality Translations Better than Human References

Keonwoong Noh [†], **Seokjin Oh**[†], and **Woohwan Jung**
Department of Applied Artificial Intelligence, Hanyang University
{rohgw011, seokjinoh, whjung}@hanyang.ac.kr

Abstract

In Machine Translation (MT) evaluations, the conventional approach is to compare a translated sentence against its human-created reference sentence. MT metrics provide an absolute score (e.g., from 0 to 1) to a candidate sentence based on the similarity with the reference sentence. Thus, existing MT metrics give the maximum score to the reference sentence. However, this approach overlooks the potential for a candidate sentence to exceed the reference sentence in terms of quality. In particular, recent advancements in Large Language Models (LLMs) have highlighted this issue, as LLM-generated sentences often exceed the quality of human-written sentences. To address the problem, we introduce the **Res**id**u**al score **Me**tric (ResuMe), which evaluates the relative quality between reference and candidate sentences. ResuMe assigns a positive score to candidate sentences that outperform their reference sentences, and a negative score when they fall short. By adding the residual scores from ResuMe to the absolute scores from MT metrics, it can be possible to allocate higher scores to candidate sentences than what reference sentences are received from MT metrics. Experimental results demonstrate that ResuMe enhances the alignments between MT metrics and human judgments both at the segment-level and the system-level.

1 Introduction

Evaluation metrics for Machine Translation (MT) mainly rely on comparisons with human-crafted reference sentences. Most frequently used metrics count the number of matching tokens between the reference and candidate sentences (Papineni et al., 2002; Popović, 2015) or estimate the similarity within the embedding space (Zhang et al., 2020). These metrics impose penalties on candidate sentences that show less alignment with the reference sentence while rewarding those that closely match the reference.

However, the current design of MT metrics leads to a limitation: they overlook the possibility that candidate sentences outperform reference sentences in terms of quality. Candidate sentences can be better than their reference sentences in reality for two reasons. Firstly, reference sentences, subject to the variability of human translators, may contain errors or inconsistencies. Thus, reference sentences are not guaranteed to be of best quality and often include critical translation and grammatical errors (Freitag et al., 2023). Secondly, Large Language Models (LLMs) exhibit remarkable generative capability. As recent advancements focus on training LLMs to follow human feedbacks (Ouyang et al., 2022), LLMs align with human preferences increasingly. Consequently, LLMs show strong performance across various generative tasks and even surpass average human performance in academic exams originally designed for humans (Achiam et al., 2023). Notably, some LLMs have been shown to produce better translations than gold human references according to automatic evaluation metrics (Xu et al., 2024) and can be rated higher than a human reference system for specific language pairs (Freitag et al., 2023). Nevertheless, conventional MT metrics inherently assign the highest scores to candidate sentences that are identical to the reference sentence, thereby suffering from the reference-bias problem (Fomicheva and Specia, 2016). As a result, current MT metrics do not allocate a higher score to a candidate sentence than the scores assigned to the reference sentence even when the candidate sentence is superior to its reference sentence. This situation requires a metric design that recognizes the potential superiority of machine-generated content.

To address the problem mentioned above, we present **Res**id**u**al score **Me**tric (ResuMe), a novel approach for MT evaluation. While traditional MT

[†] Major in Bio Artificial Intelligence

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 5111–5127
November 12-16, 2024 ©2024 Association for Computational Linguistics

metrics allocate an absolute score to the candidate sentence based on the similarity with the reference sentence, our approach RESUME quantifies the relative quality of the candidate sentence compared to the reference sentence. The relative score is positive when the candidate sentence is better than the reference sentence, and negative when the reference sentence surpasses the candidate sentence. Then, we add the relative score to the absolute score computed by an existing MT metric. By introducing RESUME, it has become possible to assign a higher score to the candidate sentence better than to the reference sentence itself. As a result, the proposed method can be utilized to appropriately evaluate LLMs that provide translation results superior to those created by humans. In the experimental section, we report an interesting observation that GPT-4 outperforms conventional MT systems with the proposed method RESUME.

Labeling a substantial amount of residual scores to train our model is both time-consuming and costly. For this reason, we propose a method to train our RESUME model using existing labeled data with standard absolute scores, eliminating the additional labeling costs. We conduct extensive experiments to validate the performance of RESUME both with the segment-level and the system-level evaluations. Specifically, RESUME encourages the alignment between MT metrics and human expert ratings across various language pairs at the segment level evaluation even without residual score labels. In system-level evaluation, our approach affords more accurate assessments for MT systems especially for LLMs which are underestimated by conventional MT metrics. Furthermore, RESUME can assess candidates outperforming references accurately compared to other MT metrics. The code is available via a GitHub repository[1]. Our main contributions are summarized as follows:

- We identify a new issue that current MT metrics fail to assign elevated scores to better candidates than references compared to the scores given to the references, suffering from the reference bias problem.
- We propose a novel MT metric, RESUME, which calculates the relative quality between reference and candidate sentences, instead of considering the reference as a perfect gold standard.
- We propose a method to generate residual (relative) scores for training RESUME by converting absolute scores of existing human-rated datasets, without any additional labeling cost.
- The empirical results demonstrate that RESUME enhances the correlation of MT metrics with human expert ratings and addresses the reference-bias problem.

2 Related Work

Lexical overlap metrics. Conventional MT metrics focus on evaluating candidate sentences at the surface-from level. Such metrics (Papineni et al., 2002; Banerjee and Lavie, 2005; Popović, 2015, 2017) count n-grams appearing in both reference and candidate sentences simultaneously. Nevertheless, these metrics fail to recognize semantically equivalent, but lexically diverse words, resulting in incorrect scores on paraphrased sentences. This attribute leads to a lower correlation with human judgements.

Embedding similarity metrics. Word embeddings are employed in order to capture semantic similarity between words. Specifically, since Pre-trained Language Models (PLMs) (Devlin et al., 2019; Liu et al., 2019) are learned from huge amounts of text data, its contextual embedding enables a deeper understanding of words between reference and candidate sentences. Zhang et al. (2020) extracts token embeddings corresponding to reference and candidate sentences using BERT or RoBERTa. Then, the final score is derived by calculating the cosine similarity among these embeddings.

Trainable metrics. Recent works (Shimanaka et al., 2018; Sellam et al., 2020) attempt to train metrics using a dataset containing human rating. These methods involve adding a regression layer on the top of PLMs and training it to directly predict human rating between reference and candidate sentences. In the machine translation, several studies (Rei et al., 2020a; Kocmi et al., 2022b; Wan et al., 2022) utilize the source sentence as semi-reference jointly during training since it holds the same meaning with reference sentence, playing an important role in translation quality assessment. This approach represents improved alignment with human ratings in machine translation evaluation.

Reference bias of MT metrics. The reference bias problem is raised by Fomicheva and Specia (2016); Ma et al. (2017); Freitag et al. (2020). Previous

[1] https://github.com/hyudsl/beyond_reference_resume

studies highlight the bias in conventional MT evaluation, where subpar candidates sharing a similar style with references receive high scores. With advancements in LLMs, a new issue of MT metrics has emerged, which has not been considered by the previous studies. Although LLM-based models can generate high-quality candidates that surpass the references, these candidates do not receive higher scores than those given to the references. In this work, we identify the new issue and propose a novel method to overcome the problem.

Zouhar and Bojar (2024) emphasizes the importance of reference quality in the automatic MT evaluation. Low-quality references can significantly degrade the performance of MT metrics, leading to unreliable assessments of candidates. However, obtaining high-quality references is both expensive and time-consuming. In this work, we introduce a new MT metric that measures the relative quality between the reference and candidates, allowing for an accurate evaluation regardless of the reference quality.

3 Conventional learnable MT metrics

The evaluation of MT commonly involves the use of source, reference, and candidate sentences, denoted as s, r, and c respectively. Given a triplet of (s, r, c), an MT metric is formulated as a scoring function $f : (s, r, c) \rightarrow \mathbb{R}$. A dataset for a learnable MT metric comprises a set of tuples $\{(s_i, r_i, c_i, y_i)\}_{i=1}^{N}$ with size of N, where y_i represents the human rating that assesses the quality of the candidate sentence c_i. The primary objective of a learnable metric is to generate scores that are correlated with human ratings. To achieve this, the training process for a learnable metric involves minimizing the mean squared error between the prediction of the metric and human ratings

$$\mathcal{L} = (y_i - f_\theta(s, r, c))^2$$

where f_θ is a learnable metric parameterized by θ.

4 Proposed Method

In this section, we first introduce the Residual score Metric (ResuMe) for machine translation (§4.1). Then, we discuss how to train ResuMe with residual scores (§4.2). Finally, we propose a method for training ResuMe with existing data that includes only absolute scores (§4.3).

4.1 The Residual Metric ResuMe

Current MT metrics are designed to yield the maximum score when evaluating the reference sentence itself, resulting in a lower score when assessing the candidate sentence (i.e., $f(s, r, c) \leq f(s, r, r)$). As a result, existing methods are inherently incapable of assigning a higher score to a candidate sentence even if it is *better* than the reference sentence (i.e., $f(s, r, c) \not> f(s, r, r)$).

To address this problem, we introduce a residual score ResuMe to indicate the relative quality of the candidate sentence compared to the reference sentence. The ResuMe score ranges from -1 to 1, outputting a positive value when the candidate sentence c surpasses the reference sentence r in quality, and a negative value when it is worse than the reference. We use the summation of ResuMe and an existing absolute score such as BLEU (Papineni et al., 2002) as the evaluation metric for MT (i.e., $f(s, r, c) + \lambda \cdot \text{ResuMe}(s, r, c)$). λ is a hyperparameter to adjust the importance and scale of the residual score. By adding this residual score to the output of an MT metric, ResuMe enables a *better* candidate sentence to receive a higher evaluation than the reference sentence (i.e., $f(s, r, c) + \lambda \cdot \text{ResuMe}(s, r, c) > f(s, r, r)$).

4.2 Training ResuMe with Residual Scores

ResuMe takes source, reference, and candidate sentences as input and should consider the relationships between the sentences. We use existing model architecture suitable for this process such as UniTE (Wan et al., 2022) and COMET (Rei et al., 2020a). We empirically demonstrate that regardless of the specific architecture used, ResuMe consistently enhances the performance of existing MT metrics in §6.6.

Our goal of learning ResuMe involves two key aspects: (1) to assign a positive score when a candidate sentence is *better*, a negative score otherwise and, (2) to achieve a higher correlation than those achieved by an existing MT metric without ResuMe. If there is an MT evaluation dataset with residual scores, ResuMe can be trained in a supervised manner. Given a tuple $(s, r, c, \Delta y)$, we optimize ResuMe to reduce the mean squared error:

$$\mathcal{L} = (\Delta y - \text{ResuMe}(s, r, c))^2$$

where Δy is the relative (residual) score of the candidate c compared to the reference r.

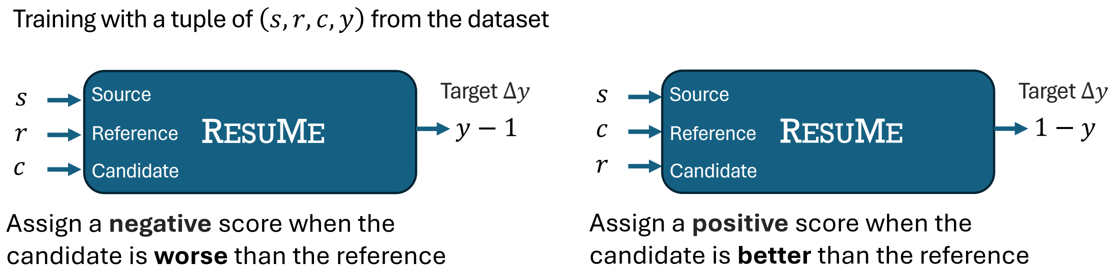
Training with a tuple of (s, r, c, y) from the dataset
s
r
c
Source
Reference
Candidate
RESUME
Target Δy
$y-1$
Assign a **negative** score when the candidate is **worse** than the reference
s
c
r
Source
Reference
Candidate
RESUME
Target Δy
$1-y$
Assign a **positive** score when the candidate is **better** than the reference

Figure 1: The training process of RESUME with absolute scores.

4.3 Training RESUME with Absolute Scores

In existing datasets for training MT metrics (Bojar et al., 2017, 2018; Barrault et al., 2019, 2020), only absolute scores for candidate sentences are available, typically ranging from 0 to 100, since our work is the first study to measure the relative score. To utilize the existing datasets, we transform these absolute scores to the score quantifying the difference between sentences. We train RESUME using the WMT 17-20 shared task data. Since the range of RESUME is from -1 to 1, we normalize the ratings in the data to a range from 0 to 1.

The straightforward method for generating residual scores is to compute the difference in absolute scores among candidate sentences that share the same source sentence. However, because multiple human raters were involved in measuring absolute scores, and their scoring strategies are varied in the existing dataset, utilizing the difference between candidate sentences can be inappropriate (Graham et al., 2013; Sellam et al., 2020; Kocmi et al., 2022b). To solve the problem, we use reference sentences by assigning a maximum score of 1 to all reference sentences in the dataset. Despite the presence of some noisy reference sentences, this approach is generally effective for learning residual scores, as the training data from WMT 17-20 data lacks translations generated by the recent LLMs, resulting in the majority of reference sentences have higher quality than their candidate sentences.

Figure 1 shows the training process of RESUME. Given a tuple (s, r, c, y), where $y \in [0, 1]$, we can set the residual score $\Delta y = \text{score}(c) - \text{score}(r) = y - 1$ by the above assumption. RESUME is trained to predict the residual score Δy by minimizing the following MSE loss

$$\mathcal{L}_{\text{ref>cand}} = ((y-1) - \text{RESUME}(s, r, c))^2.$$

However, training solely through the above process results in RESUME only returning negative values, thereby failing to allocate additional scores to the candidate sentence better than the reference.

Thus, we propose using the reference sentence r as the candidate and the candidate sentence c as the reference conversely to the previous process. In that case, the residual score is $\Delta y = \text{score}(r) - \text{score}(c) = 1 - y$ which is a positive value. Then, we enable RESUME to output a positive score to the case when the higher quality candidate sentences are assessed, using the following objective function:

$$\mathcal{L}_{\text{ref<cand}} = ((1-y) - \text{RESUME}(s, c, r))^2$$

Consequently, RESUME is trained to output a positive value when the candidate c is better than the reference r, and a negative value when it is worse than the reference.

5 Experimental Settings

5.1 Baseline Methods

We classify the baseline metrics based on whether they are trained or not on human ratings. For unsupervised metrics, we select BLEU (Post, 2018), chrF++ (Popović, 2017), and BERTScore (Zhang et al., 2020). We employ BLEURT (Sellam et al., 2020), COMET (Rei et al., 2020a), UniTE (Wan et al., 2022), MS-COMET-22 (Kocmi et al., 2022b), and xCOMET-XL (Guerreiro et al., 2023) as supervised metric. The implementation details for each metric are as follows:

- **BLEU, chrF++**: We use BLEU and chrF++ implemented in SacreBLEU (Post, 2018) repository[2]. We apply the tokenizer zh for Chinese and 13a for the other languages.
- **BERTScore**: We use XLM-RoBERTa$_{\text{LARGE}}$ version of BERTScore.
- **BLEURT**: We use the recommended checkpoint BLEURT-20[3] which is based on 32-layer

[2] https://github.com/mjpost/sacrebleu
[3] https://github.com/google-research/bleurt

RemBERT (Chung et al., 2020). It is trained on WMT 15-19 Metric Shared Task dataset and synthetic data derived from WMT corpus.

- **COMET**: We utilize the `wmt20-comet-da` checkpoint implemented in COMET framework repository[4]. It is fine-tuned on WMT 17-19 Metrics Shared Task dataset and employs XLM-RoBERTa$_{\text{LARGE}}$ as PLM.
- **UniTE**: We employ `UniTE-MUP-large` which utilizes XLM-RoBERTa$_{\text{LARGE}}$ as PLM and is fine-tuned on pseudo-labeled synthetic data and WMT 17-19 Metric Shared Task dataset.
- **MS-COMET-22**: We download the checkpoint from MS-COMET-22 repository[5]. MS-COMET-22 is fine-tuned on both WMT 17-19 Metrics Shared Task dataset and proprietary data collected from experts. MS-COMET-22 also utilizes XLM-RoBERTa$_{\text{LARGE}}$ as PLM.
- **xCOMET-XL**: We use the `XCOMET-XL` checkpoint implemented in the COMET framework repository. xCOMET-XL has 3.5B parameters finetuned on the WMT Metrics Shared Task datasets and proprietary data.

5.2 Datasets

WMT dataset. The WMT Metrics Shared Task datasets are widely used human-rated datasets for training MT metrics. The datasets employ two distinct forms of annotations for human ratings, namely Direct Assessments (DA) and Multidimensional Quality Metrics (MQM). DA ratings are derived from crowd-sourced evaluations, with scores ranging from 0 to 100. Conversely, MQM ratings are determined by experts, using a comprehensive set of error annotations and guidelines (Freitag et al., 2021). As outlined in §4.3, we construct the train set for RESUME using WMT 17-20 DA datasets (Bojar et al., 2017, 2018; Barrault et al., 2019, 2020) while comprising the valid set based on WMT 21 DA dataset (Akhbardeh et al., 2021). To validate the effectiveness of RESUME, we use WMT 22 MQM dataset (Freitag et al., 2022) as the test set. Notably, the MQM dataset includes only three language pairs (en-de, en-ru, and zh-en) due to the high cost of expert labeling. There are no overlapping sentences among the train, valid, and test sets. Training examples are provided in Appendix C. For detailed dataset statistics, please refer to Appendix G.

Post-edited translation dataset. Additionally, we utilize a post-editing dataset in MT (Fomicheva et al., 2022) to analyze how MT metrics assign a score when the candidate excels the reference. The dataset comprises source sentences, their corresponding translations, and the post-edited version of these translations refined by experts.

5.3 Measures for Meta Evaluation

Segment-level evaluation. We employ Kendall's τ-like correlation coefficient to analyze the DA dataset (Ma et al., 2019). The correlation coefficient is computed as follows:

$$\tau = \frac{|\text{Concordant}| - |\text{Discordant}|}{|\text{Concordant}| + |\text{Discordant}|}$$

For candidate sentence pairs derived from the same source sentence, |Concordant| represents the number of sentence pairs where the metric assigns a higher score to the candidate sentence that also received a higher human judgement. |Discordant| is the opposite case.

System-level evaluation. We use pairwise accuracy (Kocmi et al., 2021) on the MQM dataset. For each system pair, the difference in metric scores (metricΔ) and the difference in human scores (humanΔ) are calculated. The pairwise accuracy is defined as:

$$\text{Accuracy} = \frac{|\text{sign}(\text{metric}\Delta) = \text{sign}(\text{human}\Delta)|}{|\text{all system pairs}|}$$

5.4 Implementation Details of RESUME

We train RESUME using COMET Framework (Rei et al., 2020a) using the unified input approach (Wan et al., 2022). The model consists of a XLM and feedforward layers built on top of the XLM. The infoXLM$_{\text{BASE}}$ model (Chi et al., 2021) is employed as XLM. The feedforward network comprises 3 linear layers where the output dimensions of each layer are [3072, 1024, 1], respectively. We use AdamW optimizer (Loshchilov and Hutter, 2018) to train our model and learning rates $1e{-}5$ for PLM and $3e{-}5$ for feedforward layer with a batch size of 8. The number of epochs is 5. For a third of the first epoch, we freeze the encoder model to optimize the feedforward network, following Rei et al. (2020a). We select the checkpoint which is best correlated with the valid set. The model is trained on an NVIDIA RTX3090 GPU for 25 hours.

[4] https://github.com/Unbabel/COMET

[5] https://github.com/MicrosoftTranslator/MS-Comet

When calculating the final score with RESUME, to avoid the complexity of individual adjustments, we employ a common $\lambda = 0.2$ value in our main experiments. Despite the fixed λ value, we empirically demonstrate that RESUME consistently enhances the correlation of various MT metrics with human judgments in § 6.1. We report the optimal λ value for each MT metric in Appendix B. The impact of the λ value is explored in § 6.7.

6 Experimental Results

6.1 Main Results

Table 1 shows the performance changes of baseline methods on the WMT 22 MQM dataset when integrating RESUME. First, at the segment-level, we observe that RESUME enhances the average correlation for both supervised and unsupervised metrics. Specifically, the correlation consistently increases across all metrics for non-English-targeted as well as English-targeted pairs. This indicates that RESUME can enhance alignments of MT metrics with expert scores. Second, at the system-level, RESUME also improves the average of pairwise accuracy across all metrics. Particularly, considerable performance gains are observed in unsupervised metrics. For instance, there is a larger increase of 0.08 in BERTScore, which presents comparable results to other supervised metrics such as COMET and MS-COMET-22. This suggests that RESUME contributes to a more accurate rank of systems that are underestimated by conventional measurements. Note that the performance could be improved by applying the optimal λ value for each metric. Additionally, we provide the results on medium- and low-resource language pairs in Appendix D.

6.2 Analysis on Reference Bias of MT Metrics

Using the post-editing dataset, we conduct experiments to verify that RESUME actually assigns higher scores to a candidate that is superior to the reference rather than the scores assigned to the reference itself. Post-edited translations refined by experts exhibit superior quality than pre-edited translations. Thus, we examine whether MT metrics assign higher scores to post-edited translations than to pre-edited translations when pre-edited translations are used as the references. Table 2 represents the proportion of instances where each metric assigns higher scores to post-edited translations. Existing MT metrics allocate higher scores to post-edited translations at a very low rate, regardless of whether they are trained or not. For example, widely-used MT metrics such as COMET and UniTE exhibit rates below 7%. This demonstrates that existing MT metrics fail to evaluate candidates of superior quality than the reference correctly. However, our RESUME significantly outperforms other MT metrics, achieving a much higher rate of 59%. For the comparison with the reference-free metric, please refer to Appendix A.

6.3 Case Study

We aim to investigate whether RESUME assigns a positive score when the quality of the candidate sentence surpasses the reference sentence utilizing the post-editing dataset. In the post-editing dataset, we use pre-edited translations containing errors such as omission of words as the reference, and post-edited translations, refined by experts, as candidates. Table 3 shows the scoring results of MT metrics. red indicates the error locations in the pre-edited translations while yellow shows corrected parts in the post-edited translations.

The first example represents a case where pre-edited translations contain mistranslation errors. For BERTScore and MS-COMET-22, these metrics fail to award higher scores to the post-edited translation compared to the score of the pre-edited translation. However, RESUME allocates a positive score to the post-edited translation, enabling it to be evaluated higher than its original scores.

The second example showcases a scenario where the source's words are omitted in a pre-edited translation in addition to a mistranslation error. Due to the lexical difference from the pre-edited translation, the post-edited translation receives lower scores from BERTScore and MS-COMET-22 than scores of the pre-edited translation, despite of their higher quality. Nevertheless, RESUME assigns additional positive scores to post-edited translations, allowing them to receive a more accurate evaluation. For more examples, please refer to Appendix F.

6.4 Comparison with Reference-free Metric

Reference-free metrics estimate the quality of candidate sentences exclusively based on source sentences. By computing scores without reference sentences, reference-free metrics can avoid the issues associated with flawed reference sentences during the evaluation process. Table 4 shows the average correlation of COMET-QE (Rei et al., 2020b), which is the reference-free version of COMET,

Metrics		Segment-level (τ)				System-level (Acc.)			
		en-de	en-ru	zh-en	Avg.	en-de	en-ru	zh-en	Avg.
Supervised	xCOMET-XL	0.242	0.365	0.313	0.307	0.724	0.914	0.905	0.848
	+ RESUME	0.350	0.423	0.327	**0.367**	0.771	0.914	0.886	**0.857**
	UniTE	0.328	0.384	0.324	0.345	0.695	0.886	0.838	0.806
	+ RESUME	0.343	0.397	0.328	**0.356**	0.695	0.876	0.857	**0.810**
	BLEURT	0.330	0.352	0.304	0.329	0.781	0.886	0.867	0.844
	+ RESUME	0.361	0.384	0.309	**0.352**	0.790	0.877	0.876	**0.848**
	COMET	0.309	0.349	0.277	0.312	0.800	0.857	0.838	0.832
	+ RESUME	0.330	0.366	0.284	**0.326**	0.810	0.857	0.838	**0.835**
	MS-COMET-22	0.247	0.313	0.241	0.267	0.752	0.838	0.867	0.819
	+ RESUME	0.288	0.344	0.264	**0.299**	0.800	0.829	0.848	**0.825**
Unsupervised	BERTScore	0.231	0.287	0.228	0.248	0.714	0.867	0.714	0.765
	+ RESUME	0.316	0.348	0.258	**0.307**	0.848	0.857	0.819	**0.841**
	chrF++	0.191	0.234	0.180	0.202	0.714	0.876	0.705	0.765
	+ RESUME	0.257	0.296	0.222	**0.259**	0.762	0.905	0.733	**0.800**
	BLEU	0.119	0.122	0.110	0.117	0.667	0.838	0.695	0.733
	+ RESUME	0.241	0.264	0.192	**0.232**	0.714	0.905	0.724	**0.781**

Table 1: Results on the WMT22 MQM dataset in both segment-level and system-level. We report Kendall's τ-like correlation at segment-level, and pairwise accuracy at system-level. Values marked in **bold** signify an increase when RESUME is applied. Avg. denotes average correlation or pairwise accuracy across all language pairs.

Metrics	Pre-edited. < Post-edited.
COMET	1%
BLEURT	2%
BERTScore	3%
MS-COMET-22	4%
UniTE	7%
RESUME	**59%**

Table 2: The percentage of instances where the MT metric assigns a higher score to the post-edited translation compared to the pre-edited translation.

on WMT 22 MQM dataset. The results represent that COMET-QE exhibits a declined correlation in comparison to COMET. This observation implies that the absence of reference sentences might limit the opportunity to thoroughly assess candidate sentences with most of reliable reference sentences, leading to less accurate evaluations.

6.5 Ranking Top Performing MT Systems

Following Hendy et al. (2023), we compare GPT models and the best-performing MT system in the WMT22 en-zh pair. Figure 2 represents the scoring results when RESUME is either applied or not. The top-ranked system in WMT22 en-zh pair is marked in 'Online-W' (Kocmi et al., 2022a). In contrast to the human evaluation in Hendy et al. (2023), BLEU selects the 'Online-W' system as the highest-performing system in the en-zh pair. Nevertheless, the incorporation of RESUME into the evaluation process aligns with the previous finding, showcasing GPT-3's superiority over the Online-W system. Additionally, the summation of RESUME with conventional MT metric produces an interesting result by choosing GPT-4, known for its superior generative performance (Achiam et al., 2023), as the best MT system in the en-zh pair. These results indicate that RESUME precisely evaluates high-performance MT systems.

6.6 Architecture of RESUME

We investigate whether the performance enhancement of MT metrics is due to the specific model architecture employed for RESUME. Figure 3 shows the variations in the average correlation of MT metrics on WMT 22 MQM dataset at the segment-level when varying the model architecture of RESUME. 'RESUME (UniTE)' denotes the model trained using the UniTE architecture while 'RESUME (COMET)' indicates the mode trained using the COMET architecture. A consistent trend is ob-

Example #1						
Source	Кошка любит молоко, да рыло коротко.					
Pre-edited.	The cat loves milk, but the fish is short.					
Post-edited.	The cat loves milk, but the snout is too short.					
Scores	**Reference: `Pre` , Candidate: `Pre`**		**Reference: `Pre` , Candidate: `Post`**			
	BERTScore	0.997	BERTScore	0.965	BERTScore + RESUME	1.099 (= 0.965 + 0.134)
	MS-COMET-22	0.868	MS-COMET-22	0.818	MS-COMET-22 + RESUME	0.952 (= 0.818 + 0.134)
Example #2						
Source	Бить по рогам лишь надсада рукам .					
Pre-edited.	Beating the horns is only a handshake .					
Post-edited.	Beating the horns only hurts your hands .					
Scores	**Reference: `Pre` , Candidate: `Pre`**		**Reference: `Pre` , Candidate: `Post`**			
	BERTScore	0.999	BERTScore	0.953	BERTScore + RESUME	1.118 (0.953 + 0.165)
	MS-COMET-22	0.877	MS-COMET-22	0.741	MS-COMET-22 + RESUME	0.906 (0.691 + 0.165)

Table 3: Two examples from the post-editing translation dataset. **`Pre`** and **`Post`** denote pre-edited translations and post-edited translations, respectively.

Metrics	**Avg. τ**
COMET-QE	0.236
COMET	0.312
+ RESUME	**0.326**

Table 4: Comparison of the average correlation among MT metrics with and without reference utilization.

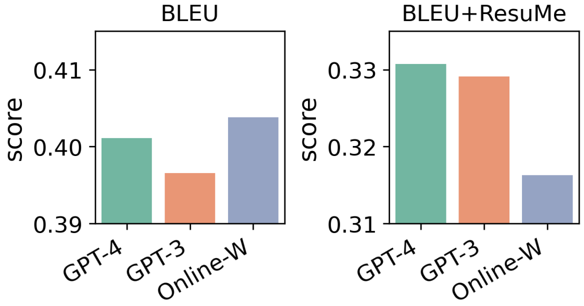

Figure 2: Scoring results on GPTs and top-performing system (Online-W) in WMT22 en-zh pair.

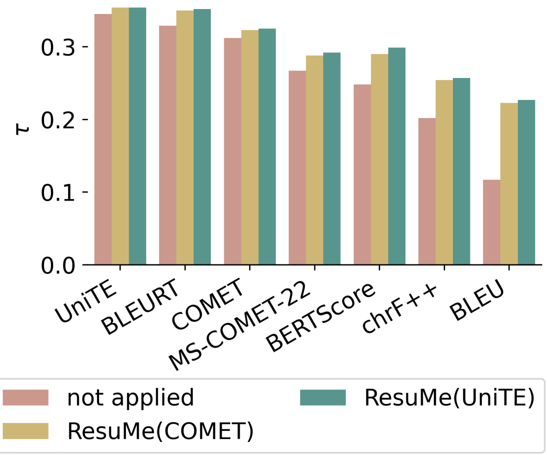

Figure 3: The impact of model architecture for RESUME on the average correlation of MT metrics.

served across all MT metrics: the application of RESUME improves the correlation, regardless of the specific architecture. Concretely, 'RESUME (UniTE)' increases the correlation of MT metrics more than 'RESUME (COMET)' across all metrics. Hence, we select UniTE as the backbone architecture of RESUME.

6.7 Impacts of Residual Scores Ratio

We conduct an experiment to examine the degree to which the integration of RESUME scores contributes to the enhancements of MT metrics performance. To do this, we vary the weight λ from 0.1 to 10 to investigate the importance of residual score. Figure 4 represents the changes in the average correlation of MT metrics on WMT 22 MQM dataset at the segment-level. Depending on the existing metric, the optimal values of λ are different. Specifically, many supervised metrics attain higher correlation when $\lambda < 0.5$, while most unsupervised metrics attain near $\lambda = 1$. Thus, we recommend using the weight $\lambda \leq 1$ for existing MT metrics. In our experiment, we use $\lambda = 0.2$ for all metrics instead of using individual optimal values to avoid overfitting.

7 Conclusion

In this paper, we propose RESUME, a novel metric designed to assess the relative quality between reference and candidate sentences in MT. By adding residual scores from RESUME to the output of ex-

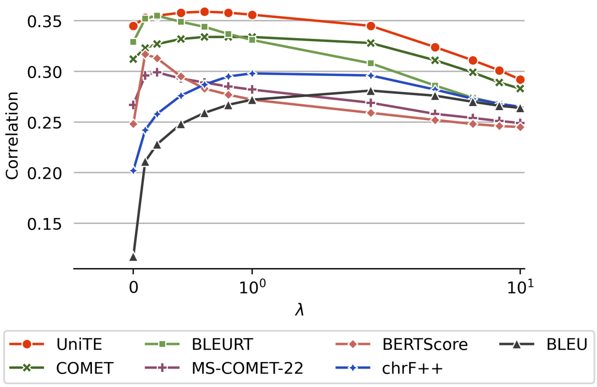

Figure 4: Changes in the average correlation of MT metrics according to RESUME score ratios.

isting MT metrics, we address the reference-bias problem in current MT evaluation, which arises with significant advancements in MT systems: the candidate sentences, produced by MT systems, excel the quality of the reference sentences. The empirical results demonstrate that RESUME not only increases the alignment between MT metrics and human judgements but also enables to discriminate between top MT systems. Furthermore, we disclose analyses that RESUME indeed evaluates higher quality translations over references more accurately against other MT metrics.

Limitations

Although RESUME improves the performance of existing MT metrics, the evaluation time may be slightly longer because the final scores are obtained by adding the residual scores with the output of other MT metrics. One of the methods generating residual scores could calculate the scores of candidate sentences in existing human-rated datasets, as mentioned in §4.3. However, when we trained RESUME containing this method, we observed the decreased performance of RESUME.

Despite the good performance of RESUME in our experiments, we partly agree that the assumption of assigning a reference score of 1 during the training process is revisited in future works to fully leverage recent datasets (e.g., WMT 22, 23) that include a growing proportion of LLM-generated translations. In addition, we believe that RESUME could be enhanced by including human-annotated residual scores into the training data.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.RS-2023-00261068, Development of Lightweight Multimodal Anti-Phishing Models and Split-Learning Techniques for Privacy-Preserving Anti-Phishing). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-00155885, Artificial Intelligence Convergence Innovation Human Resources Development (Hanyang University ERICA)).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. 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.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos

Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2020. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Marina Fomicheva and Lucia Specia. 2016. Reference bias in monolingual machine translation evaluation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 77–82, Berlin, Germany. Association for Computational Linguistics.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974, Marseille, France. European Language Resources Association.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. 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, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. 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.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2023. xcomet: Transparent machine translation evaluation through fine-grained error detection. *arXiv preprint arXiv:2310.10482*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan

Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

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. 2022a. 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, 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.

Tom Kocmi, Hitokazu Matsushita, and Christian Federmann. 2022b. MS-COMET: More and better human judgements improve metric performance. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 541–548, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Qingsong Ma, Yvette Graham, Timothy Baldwin, and Qun Liu. 2017. Further investigation into reference bias in monolingual evaluation of machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2476–2485, Copenhagen, Denmark. Association for Computational Linguistics.

Qingsong Ma, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Florence, Italy. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020a. 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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020b. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. 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.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 751–758, Belgium, Brussels. Association for Computational Linguistics.

Yu Wan, Dayiheng Liu, Baosong Yang, Haibo Zhang, Boxing Chen, Derek Wong, and Lidia Chao. 2022. UniTE: Unified translation evaluation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8117–8127, Dublin, Ireland. Association for Computational Linguistics.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *Forty-first International Conference on Machine Learning*.

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*.

Vilém Zouhar and Ondřej Bojar. 2024. Quality and quantity of machine translation references for automatic metrics. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 1–11, Torino, Italia. ELRA and ICCL.

A Comparison with Reference-free Metrics on a Post-editing Dataset

Metrics	**Pre-edited.** **< Post-edited.**
COMET	1%
COMET-QE	54%
ResuMe	**59%**

Table 5: The percentage of instances where the MT metrics assign a higher score to the post-edited translation compared to the pre-edited translation.

We investigate that reference-free metric assigns higher scores to post-edited translations when using pre-edited translations as references, as outlined in § 6.2. Table 5 shows the result of COMET-QE, the reference-free version of COMET, on the post-editing dataset. COMET-QE allocates greater scores to post-edited translations more frequently, compared to COMET, as it is not influenced by the quality of references. However, COMET-QE shows a lower rate compared to our ResuMe. This indicates that our ResuMe evaluates higher-quality translations more accurately than the method without references.

B The Optimal Ratio of ResuMe Score for MT Metrics on the WMT Dataset

We investigate the most suitable proportion (λ) of ResuMe scores for MT metrics on WMT dataset. We diversify the ratio λ from 0.1 to 10 incrementally. Table 6 shows the optimal λ value for MT metrics at both segment- and system-level.

C Training Examples used for ResuMe

To facilitate understanding of the training process of ResuMe, we provide training examples with residual score in Table 8. The original data example with the absolute score is represented in Table 7. As introduced in § 4.3, we convert absolute scores from a human-rated dataset into residual scores without any additional labeling cost.

Metrics	**Segment-level**	**System-level**
UniTE	0.4	0.6
BLEURT	0.2	0.2
COMET	0.6	0.6
MS-COMET-22	0.2	0.2
BERTScore	0.2	0.2
chrF++	1	1
BLEU	2	1

Table 6: The best-suited ratio(λ) of ResuMe score for MT metrics on the WMT 22 MQM dataset.

D Results on Medium- and Low-Resource Language Pairs

Since there are only three high-resource language pairs in the WMT 22 MQM dataset, we additionally test ResuMe on 11 medium- and low-resource language pairs in the WMT 22 DA dataset. As shown in Table 9 and 10, ResuMe consistently improves the average performance of baseline methods in medium- and low-resource language pairs.

E Results on the WMT 21, 22, and 23 MQM tasks.

To validate the generalization performance of ResuMe, we experiment with the WMT 20, 21, and 23 MQM datasets. To avoid overlap between the training and evaluation datasets, we utilize ResuMe trained on the WMT 17-18 DA datasets for the WMT 20 MQM dataset and ResuMe trained on the WMT 17-19 DA datasets for the WMT 21 MQM dataset. For the WMT 23 MQM dataset, we employ the same ResuMe checkpoint used in our main experiment without any additional training. As shown in Table 11, 12, and 13, ResuMe enhances the performance of baseline metrics in the newly introduced dataset at both segment- and system-levels.

F Case Study

We sample several (source, reference) pairs from the WMT 21 de-en language pair where reference sentences represent deficiencies, such as the omission of words. Subsequently, we generate candidate sentences using GPT-4, as it can produce high-quality translations than gold references (Xu et al.,

Source	Reference	Candidate	Absolute Score
What is your view of the match?	Jak byste zhodnotily zápas?	Jaký je váš pohled na zápas?	0.892

Table 7: An example sampled from an existing human-rated dataset with absolute scores.

Source	Reference	Candidate	Residual Score
What is your view of the match?	Jak byste zhodnotily zápas?	Jaký je váš pohled na zápas?	-0.108 (= 0.892 - 1)
What is your view of the match?	Jaký je váš pohled na zápas?	Jak byste zhodnotily zápas?	0.108 (= 1 - 0.892)

Table 8: Training examples used for RESUME.

2024). These candidate sentences are then scored. Table 14 shows the scoring results assigned by existing MT metrics and RESUME.

The first example presents that the English word corresponding to 'In Bangladesch' does not exist in reference. In the case of BERTScore and MS-COMET-22, these metrics do not assign higher scores to the candidate sentence rather than the evaluation of the reference sentence itself. However, RESUME gives a positive value to the candidate sentence, allowing it to receive a higher assessment.

The second example shows a case where certain words, found in the reference, are absent from the source. Notably, there are no English counterparts for 'Würzburg' and 'echzig' in the reference. Moreover, the 'the 1860s' word appears in the reference sentence but does not exist in the source sentence. The assessment of the candidate is decreased using BERTScore and MS-COMET-22. However, RESUME assigns an additional score, resulting in proper evaluation for the candidate sentence.

Furthermore, we also provide examples from the post-editing dataset. As shown in Table 15, existing MT metrics struggle to assign higher scores to post-edited translations than to erroneous pre-edited translations when pre-edited translations serve as references. Nevertheless, since RESUME can assess translations regardless of the reference quality, post-edited translations can receive higher evaluations, when RESUME is applied, despite the presence of flawed references.

G Dataset Statistics

We provide detailed statistics on the datasets utilized for training and evaluating RESUME. Table 16 shows the count of sentence pairs for each language pair in the WMT Metrics Shared Task datasets.

Metrics		en-ja	cs-uk	en-hr	en-uk	sah-ru	uk-cs	en-cs	ja-en	cs-en	uk-en	liv-en	Avg.
Supervised	UniTE	0.108	0.195	0.184	0.141	0.643	0.226	0.090	0.078	0.035	-0.028	0.515	0.199
	+ ResuMe	0.119	0.206	0.208	0.159	0.647	0.239	0.124	0.084	0.049	0.001	0.519	**0.214**
	BLEURT	0.068	0.149	0.113	0.078	0.632	0.159	0.033	0.042	-0.012	-0.133	0.512	0.149
	+ ResuMe	0.103	0.187	0.189	0.146	0.639	0.216	0.116	0.087	0.045	-0.018	0.527	**0.203**
	COMET	0.102	0.165	0.167	0.122	0.644	0.184	0.053	0.078	0.037	-0.040	0.523	0.185
	+ ResuMe	0.114	0.191	0.199	0.146	0.648	0.221	0.092	0.091	0.058	0.000	0.527	**0.208**
	MS-COMET-22	0.079	0.155	0.127	0.101	0.587	0.169	0.071	0.073	-0.019	-0.089	0.490	0.159
	+ ResuMe	0.104	0.175	0.178	0.144	0.591	0.204	0.140	0.097	0.015	-0.013	0.511	**0.195**
Unsupervised	BERTScore	0.060	0.114	0.087	0.041	0.628	0.115	-0.019	0.042	-0.008	-0.132	0.523	0.132
	+ ResuMe	0.099	0.154	0.176	0.125	0.617	0.188	0.121	0.086	0.028	-0.014	0.521	**0.191**
	chrF++	0.042	0.091	0.062	0.027	0.630	0.120	-0.036	0.046	-0.006	-0.144	0.481	0.119
	+ ResuMe	0.097	0.141	0.162	0.117	0.644	0.193	0.051	0.097	0.052	-0.022	0.519	**0.186**
	BLEU	0.031	0.024	-0.035	-0.047	0.580	0.047	-0.082	-0.010	-0.031	-0.185	0.423	0.065
	+ ResuMe	0.087	0.123	0.149	0.115	0.630	0.180	0.053	0.100	0.048	-0.022	0.493	**0.178**

Table 9: Segment-level Kendall's τ-like correlation results on the mid- and low-resource language pairs in the WMT 22 DA dataset. Values marked in **bold** signify an increase when ResuMe is applied. Avg. denotes the average score across all language pairs.

Metrics		en-ja	cs-uk	en-hr	en-uk	sah-ru	uk-cs	en-cs	ja-en	cs-en	uk-en	liv-en	Avg.
Supervised	UniTE	0.923	0.964	0.911	0.917	1.000	0.945	0.712	0.571	0.712	0.694	1.000	0.850
	+ ResuMe	0.912	0.964	0.933	0.889	1.000	0.982	0.727	0.571	0.727	0.694	1.000	**0.854**
	BLEURT	0.824	0.927	0.933	0.889	1.000	0.982	0.682	0.593	0.727	0.694	1.000	0.841
	+ ResuMe	0.846	0.909	0.956	0.889	1.000	0.982	0.773	0.571	0.727	0.694	1.000	**0.850**
	COMET	0.857	0.927	0.956	0.917	1.000	0.964	0.682	0.549	0.712	0.667	1.000	0.839
	+ ResuMe	0.868	0.945	0.978	0.917	1.000	0.964	0.682	0.560	0.727	0.667	1.000	**0.846**
	MS-COMET-22	0.890	0.873	0.956	0.833	1.000	0.964	0.697	0.593	0.727	0.667	1.000	0.836
	+ ResuMe	0.901	0.909	0.933	0.861	1.000	0.945	0.788	0.582	0.667	0.667	1.000	**0.841**
Unsupervised	BERTScore	0.879	0.873	0.822	0.861	1.000	0.891	0.606	0.637	0.712	0.722	0.933	0.812
	+ ResuMe	0.901	0.945	0.933	0.889	1.000	0.909	0.818	0.582	0.652	0.722	1.000	**0.850**
	chrF++	0.824	0.800	0.822	0.889	1.000	0.945	0.591	0.615	0.712	0.750	0.933	0.807
	+ ResuMe	0.846	0.873	0.933	0.889	1.000	0.927	0.606	0.626	0.697	0.722	1.000	**0.829**
	BLEU	0.725	0.800	0.756	0.861	1.000	0.927	0.591	0.593	0.697	0.750	0.933	0.785
	+ ResuMe	0.890	0.818	0.867	0.889	1.000	0.964	0.621	0.604	0.682	0.722	1.000	**0.823**

Table 10: System-level pairwise accuracy results on the mid- and low-resource language pairs in the WMT 22 DA dataset. Values marked in **bold** signify an increase when ResuMe is applied. Avg. denotes the average score across all language pairs.

Metrics		Segment-level (τ)				System-level (Acc.)			
		en-de	zh-en	he-en	Avg.	en-de	zh-en	he-en	Avg.
Supervised	xCOMET-XL	0.537	0.448	0.396	0.460	0.897	0.842	0.949	0.896
	+ ResuMe	0.574	0.471	0.371	**0.472**	0.936	0.858	0.962	**0.919**
	UniTE	0.540	0.442	0.393	0.458	0.923	0.833	0.923	0.893
	+ ResuMe	0.545	0.450	0.391	**0.462**	0.923	0.842	0.923	**0.896**
	BLEURT	0.476	0.347	0.390	0.404	0.949	0.800	0.897	0.882
	+ ResuMe	0.513	0.387	0.398	**0.433**	0.936	0.808	0.923	**0.889**
	COMET	0.544	0.428	0.383	0.452	0.923	0.825	0.872	0.873
	+ ResuMe	0.551	0.436	0.388	**0.458**	0.923	0.833	0.872	**0.876**
	MS-COMET-22	0.425	0.273	0.326	0.341	0.897	0.850	0.936	0.894
	+ ResuMe	0.506	0.393	0.288	**0.396**	0.897	0.858	0.949	**0.901**
Unsupervised	BERTScore	0.541	0.410	0.362	0.438	0.936	0.792	0.846	0.858
	+ ResuMe	0.569	0.462	0.341	**0.457**	0.910	0.858	0.897	**0.888**
	chrF++	0.337	0.211	0.310	0.286	0.910	0.750	0.769	0.810
	+ ResuMe	0.392	0.290	0.332	**0.338**	0.910	0.775	0.782	**0.822**
	BLEU	0.375	0.195	0.256	0.275	0.936	0.767	0.795	0.833
	+ ResuMe	0.441	0.327	0.303	**0.357**	0.949	0.783	0.795	**0.842**

Table 11: Results on the WMT 23 MQM dataset.

Metrics		**Segment-level(τ)**				**System-level(Acc.)**			
		en-de	**en-ru**	**zh-en**	**Avg.**	**en-de**	**en-ru**	**zh-en**	**Avg.**
Supervised	UniTE	0.179	0.317	0.114	0.203	0.728	0.617	0.495	0.613
	+ ResuMe	0.216	0.367	0.107	**0.230**	0.853	0.750	0.590	**0.731**
	BLEURT	0.194	0.301	0.103	0.199	0.831	0.708	0.476	0.672
	+ ResuMe	0.234	0.350	0.085	**0.223**	0.875	0.825	0.610	**0.770**
	COMET	0.163	0.280	0.096	0.180	0.743	0.642	0.514	0.633
	+ ResuMe	0.213	0.340	0.093	**0.215**	0.875	0.767	0.600	**0.747**
	MS-COMET-22	0.163	0.304	0.095	0.187	0.824	0.717	0.543	0.695
	+ ResuMe	0.207	0.355	0.060	**0.207**	0.846	0.792	0.610	**0.749**
Unsupervised	BERTScore	0.075	0.192	0.046	0.104	0.632	0.567	0.476	0.558
	+ ResuMe	0.200	0.314	0.042	**0.185**	0.860	0.800	0.600	**0.753**
	chrF++	0.057	0.123	0.031	0.070	0.654	0.550	0.467	0.557
	+ ResuMe	0.170	0.300	0.063	**0.178**	0.853	0.808	0.486	**0.716**
	BLEU	0.018	0.015	0.007	0.013	0.654	0.542	0.476	0.557
	+ ResuMe	0.146	0.262	0.052	**0.153**	0.757	0.800	0.448	**0.668**

Table 12: Results on the WMT 21 MQM dataset.

Metrics		**Segment-level(τ)**			**System-level(Acc.)**		
		en-de	**zh-en**	**Avg.**	**en-de**	**zh-en**	**Avg.**
Supervised	UniTE	0.234	0.217	0.226	0.733	0.467	0.600
	+ ResuMe	0.319	0.233	**0.276**	0.911	0.778	**0.844**
	BLEURT	0.320	0.320	0.320	0.933	0.578	0.756
	+ ResuMe	0.372	0.321	**0.347**	0.889	0.800	**0.844**
	COMET	0.207	0.183	0.195	0.778	0.578	0.678
	+ ResuMe	0.316	0.206	**0.261**	0.911	0.689	**0.800**
	MS-COMET-22	0.221	0.175	0.198	0.800	0.622	0.711
	+ ResuMe	0.322	0.198	**0.260**	0.844	0.800	**0.822**
Unsupervised	BERTScore	0.092	0.103	0.097	0.511	0.400	0.456
	+ ResuMe	0.312	0.182	**0.247**	0.889	0.756	**0.822**
	chrF++	0.049	0.057	0.053	0.489	0.444	0.466
	+ ResuMe	0.255	0.164	**0.209**	0.956	0.489	**0.722**
	BLEU	0.014	0.010	0.012	0.422	0.489	0.456
	+ ResuMe	0.207	0.136	**0.171**	0.800	0.467	**0.633**

Table 13: Results on the WMT 20 MQM dataset.

Example #1				
Source	In Bangladesch bemüht sich der Rote Halbmond in Booten die isolierten Menschen zu erreichen.			
Reference	The Red Crescent strove to reach isolated people in boats.			
GPT-4	In Bangladesh, the Red Crescent is endeavoring to reach isolated individuals using boats.			
Scores	Reference itself		GPT-4	
	BERTScore	1.000	**BERTScore + ReSuME**	1.067 (0.952 + 0.115)
	MS-COMET-22	0.928	**MS-COMET-22 + ReSuME**	1.042 (0.927 + 0.115)
Example #2				
Source	Sollte Zweitliga-Aufsteiger Würzburg das Finale erreichen, wäre Sechzig bereits fix für die erste Runde im DFB-Pokal qualifiziert.			
Reference	Should the new second-league team reach the finals, then the 1860s would already qualify for the first round of the DFB-Pokal.			
GPT-4	If second division promoted team Würzburg were to reach the final, Sechzig would already be qualified for the first round of the DFB-Pokal.			
Scores	Reference itself		GPT-4	
	BERTScore	1.000	**BERTScore + ReSuME**	1.119 (0.941 + 0.178)
	MS-COMET-22	0.923	**MS-COMET-22 + ReSuME**	1.071 (0.893 + 0.178)

Table 14: Examples of noisy reference sentences. yellow indicates the tokens present in both source and GPT-4 translation, but not in the reference. Conversely, red denotes the tokens in the reference, but not in the source.

Example #1						
Source	Богатый дивится , чем голь живится.					
Pre-edited.	Rich marvel at how gold lives.					
Post-edited.	Rich marvel at how the poor live.					
Scores	Reference: Pre , Candidate: Pre		Reference: Pre , Candidate: Post			
	BERTScore	0.999	BERTScore	0.903	BERTScore + ReSUME	1.082 (= 0.903 + 0.179)
	MS-COMET-22	0.864	MS-COMET-22	0.792	MS-COMET-22 + ReSUME	0.971 (= 0.792 + 0.179)
Example #2						
Source	Что сделать чтобы прошло жжение в глазах.					
Pre-edited.	What to do to get your eyes burning.					
Post-edited.	What does one do to stop the eyes burning.					
Scores	Reference: Pre , Candidate: Pre		Reference: Pre , Candidate: Post			
	BERTScore	0.999	BERTScore	0.931	BERTScore + ReSUME	1.034 (0.931 + 0.103)
	MS-COMET-22	0.873	MS-COMET-22	0.841	MS-COMET-22 + ReSUME	0.944 (0.841 + 0.103)
Example #3						
Source	Почему вам пора перестать просто ПОДНИМАТЬ штангу и гантели?					
Pre-edited.	Why should you stop just HAPPY the barbell and dumbbells?					
Post-edited.	Why should you stop just LIFTING the barbell and dumbbells?					
Scores	Reference: Pre , Candidate: Pre		Reference: Pre , Candidate: Post			
	BERTScore	0.998	BERTScore	0.948	BERTScore + ReSUME	1.057 (0.948 + 0.109)
	MS-COMET-22	0.837	MS-COMET-22	0.811	MS-COMET-22 + ReSUME	0.920 (0.811 + 0.109)
Example #4						
Source	И потеряла надежду Мария поймать кайф .					
Pre-edited.	And Maria lost hope of catching the bug .					
Post-edited.	And Maria lost hope of getting high .					
Scores	Reference: Pre , Candidate: Pre		Reference: Pre , Candidate: Post			
	BERTScore	1.0	BERTScore	0.973	BERTScore + ReSUME	1.106 (0.973 + 0.133)
	MS-COMET-22	0.902	MS-COMET-22	0.850	MS-COMET-22 + ReSUME	0.983 (0.850 + 0.133)

Table 15: Additional examples from the post-editing translation dataset. `Pre` and `Post` denote pre-edited translations and post-edited translations, respectively.

WMT 17 (DA)		WMT 18 (DA)		WMT 19 (DA)		WMT 20 (DA)		WMT 21 (DA)		WMT 22 (MQM)	
LP	Num.	LP	Num.	LP	Num.	LP	Num.	LP	Num.	LP	Num.
en-tr	2039	en-tr	3132	fr-de	3999	ps-en	4611	zu-xh	2502	en-ru	19725
zh-en	26419	de-en	28404	en-de	75777	zh-en	55006	zh-en	36128	zh-en	28125
de-en	21704	zh-en	25352	de-en	31887	en-zh	29950	en-zh	27876	en-de	19725
en-cs	20532	en-zh	22128	zh-en	20170	ru-en	19499	de-en	18409		
ru-en	17980	et-en	20496	ru-en	19644	en-de	19410	en-ja	14856		
en-ru	17358	en-ru	16748	en-zh	19506	de-en	17188	ja-en	14482		
tr-en	17335	fi-en	14965	en-cs	18201	en-cs	14575	en-de	13572		
fi-en	15159	en-et	13376	en-ru	17417	pl-en	11816	ha-en	13171		
cs-en	11585	ru-en	13157	en-fi	16079	en-ru	11226	en-is	10838		
en-zh	10221	tr-en	12851	fi-en	16021	en-pl	10572	en-ha	10812		
en-fi	10159	en-de	10208	de-cs	13804	en-ja	9578	en-cs	10006		
en-de	7025	cs-en	8732	lt-en	10315	ja-en	8939	en-ru	9313		
en-lv	5810	en-fi	8097	gu-en	9063	en-ta	7890	ru-en	8926		
en-tr	2039	en-cs	7629	en-lt	8959	ta-en	7577	is-en	8903		
		en-tr	3132	en-kk	8219	cs-en	7530	de-fr	8258		
				en-gu	6924	km-en	4722	cs-en	7242		
				kk-en	6789	ps-en	4611	fr-de	7196		
				de-fr	6691			hi-bn	4512		
				fr-de	3999			bn-hi	4461		
								xh-zu	2968		

Table 16: The number of sentence pairs in the WMT Metrics Shared Task datasets. LP denotes the language pair.