Test-time Adaptation for Machine Translation Evaluation by Uncertainty Minimization

Runzhe Zhan[1] **Xuebo Liu**[2*] **Derek F. Wong**[1*] **Cuilian Zhang**[1]
Lidia S. Chao[1] **Min Zhang**[2]
[1]NLP[2]CT Lab, Department of Computer and Information Science, University of Macau
nlp2ct.{runzhe, cuilian}@gmail.com, {derekfw, lidiasc}@um.edu.mo
[2]Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen, China
{liuxuebo, zhangmin2021}@hit.edu.cn

Abstract

The neural metrics recently received considerable attention from the research community in the automatic evaluation of machine translation. Unlike text-based metrics that have interpretable and consistent evaluation mechanisms for various data sources, the reliability of neural metrics in assessing out-of-distribution data remains a concern due to the disparity between training data and real-world data. This paper aims to address the inference bias of neural metrics through uncertainty minimization during test time, without requiring additional data. Our proposed method comprises three steps: uncertainty estimation, test-time adaptation, and inference. Specifically, the model employs the prediction uncertainty of the current data as a signal to update a small fraction of parameters during test time and subsequently refine the prediction through optimization. To validate our approach, we apply the proposed method to three representative models and conduct experiments on the WMT21 benchmarks. The results obtained from both in-domain and out-of-distribution evaluations consistently demonstrate improvements in correlation performance across different models. Furthermore, we provide evidence that the proposed method effectively reduces model uncertainty. The code is publicly available at https://github.com/NLP2CT/TaU.

1 Introduction

The evaluation of machine translation (MT) systems aims to quantitatively assess their performance using either automatic metrics or human evaluators. When developing cutting-edge MT systems, selecting the optimal model using automatic metrics is highly significant to save human labor, given a large number of candidate models. Over the last decade, the researchers have primarily relied on traditional metrics based on text overlap (Papineni et al., 2002; Snover et al., 2006; Popović, 2015) to evaluate system performance. However, these metrics fall short in capturing semantic-level information and exhibit poor correlation with human ratings when assessing the latest neural MT systems because of increased model capacity (Ma et al., 2019; Mathur et al., 2020). Consequently, several neural metrics (Zhang et al., 2020; Rei et al., 2020; Sellam et al., 2020; Zhan et al., 2021a; Wan et al., 2022) and test sets (Müller et al., 2018; Stanovsky et al., 2019; Zhan et al., 2021b; Freitag et al., 2021b) have been proposed to provide broader evaluation perspectives and show outstanding performance in evaluating state-of-the-art systems. Despite the superiority of neural metrics, the adoption of these metrics over traditional overlap-based measures has witnessed a gradual pace. The people engaged in MT research and industry remain cautious due to concerns surrounding potential robustness issues, thereby hindering the progress of popularizing neural metrics.

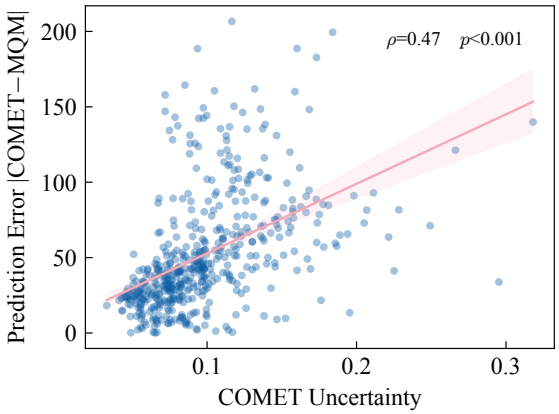


Figure 1: Correlation plot of model uncertainty of COMET with prediction error when evaluating an English-Russian MT system from WMT21 submission. The prediction error is the difference between COMET score and MQM (Multidimensional Quality Metrics; Freitag et al., 2021a) expert-based human evaluation score. COMET scores are scaled up to MQM range.

*Co-corresponding author

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 807–820

The source of robustness problem can be attributed to data shift. The fine-tuning data used when developing neural metrics is composed of labels derived from human ratings obtained when evaluating strong MT systems in the News domain, which largely limits the generalization capability of the obtained model. In real-world scenarios, the evaluation metric must be capable of assessing text originating from diverse domains with varying levels of quality. However, neural metrics, trained on limited data, may exhibit biases when dealing with out-of-distribution data. These factors present challenges in establishing neural metrics as reliable evaluation measures across a wide range of applications. Glushkova et al. (2021) proposed employing uncertainty quantification (Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017) to assess the risk associated with utilizing neural metrics in evaluation and discovered a correlation between model uncertainty and model prediction errors, as depicted in Figure 1. While Glushkova et al. (2021) have explored the uncertainty of neural metrics, the quest for a solution to mitigate uncertainty in MT evaluation remains an under-explored research area. One intuitive approach is fine-tuning the model using diverse and multi-domain data. Unfortunately, there is currently no publicly available dataset that satisfies this requirement.

In this paper, we propose an unsupervised approach for neural metrics aimed at minimizing uncertainty during test time and mitigating the challenges posed by out-of-distribution data. Our proposed method involves two additional stages integrated before the normal inference process: uncertainty estimation and test-time adaptation. Firstly, our model leverages the Monte Carlo approach (Gal and Ghahramani, 2016) to estimate the uncertainty of the current input data. Subsequently, the estimated uncertainty serves as a guiding signal to optimize a small fraction of model parameters using gradient descent. Finally, the model proceeds with the regular inference procedure, utilizing the adapted parameters to make predictions. In this way, the model can adjust its parameters dynamically to better cope with diverse data, which is flexible and does not require any labeled data.

We use the representative metric family COMET (Rei et al., 2020) as our testbed and conduct experiments on WMT21 benchmark (Freitag et al., 2021b), which accounts for evaluating out-of-distribution data. The experimental results show that our method can improve the system-level correlation performance as well as the ranking accuracy of partial COMET baselines. Furthermore, our analysis highlights the applicability of our method and confirms its efficacy in reducing uncertainty.

2 Background

MT Metrics Ideally, human labor is used to evaluate the translation quality of MT models and identify the optimal model. Since human assessment is expensive, there is a need for automatic evaluation methods that can provide instantaneous measurements of a model's capability. More specifically, given the model hypothesis h, ground truth r, and source s, the metric $\mathrm{M}(\cdot)$ will quantify the translation quality q by comparing the model hypothesis and reference $\langle h, t \rangle$:

$$q = \begin{cases} \mathrm{M}(h, s) & s = \emptyset \\ \mathrm{M}(h, r) & t = \emptyset \\ \mathrm{M}(h, s, r) & s, t \neq \emptyset \end{cases} \quad (1)$$

There are three types of metrics based on their utilization of reference information: reference-based metric $\mathrm{M}(\langle h, \cdot, r \rangle)$ (which solely utilize the target translation or jointly consider both the source and target information), and reference-free metric $\mathrm{M}(\langle h, s \rangle)$ (which solely rely on the source input). Among these, reference-based metrics are widely employed, and reference-free metrics are often categorized as quality estimation metrics (Fonseca et al., 2019).

The neural metrics build a regression scoring model by leveraging pre-trained representation, which have achieved remarkable performance in MT evaluation. In this way, the metric M is parameterized by model θ:

$$q = \mathrm{M}(\langle h, s, r \rangle; \theta) \quad (2)$$

As an example, the COMET (Rei et al., 2020) framework employs two distinct downstream architectures to leverage a pre-trained XLM (Conneau et al., 2020) model. It fine-tunes the additional regression and ranking models using human rating data obtained from the WMT Metrics task, ensuring that the tuned parameters can evaluate the translation quality.

Uncertainty As deep neural networks are widely used in real-world applications, uncertainty is a critical measurement that indicates how a model is confident in the predictions in order to prevent causing

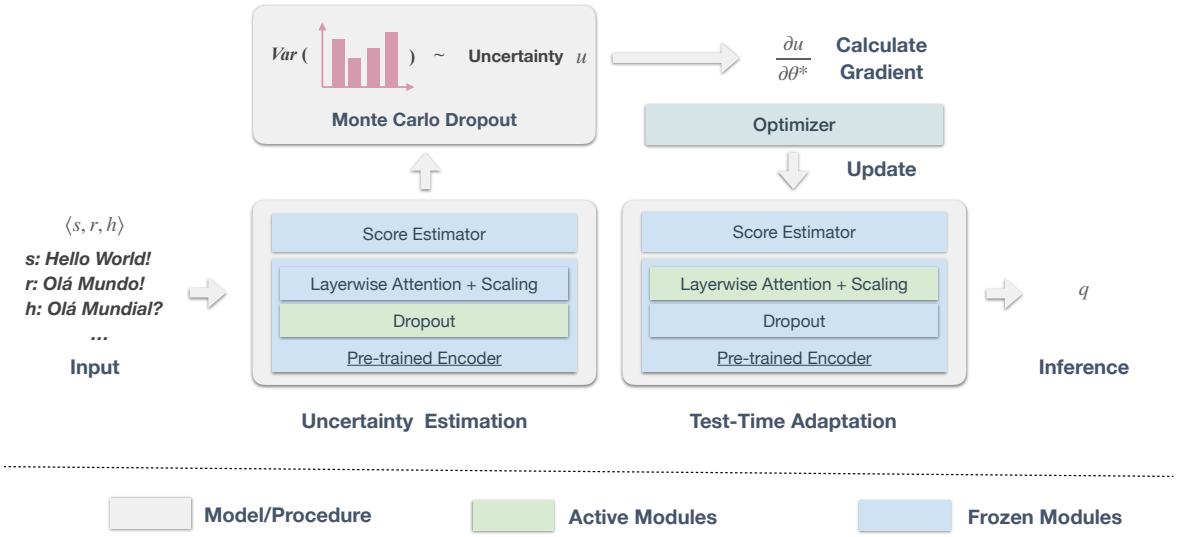


Figure 2: Illustration of the proposed method: **T**est-time **A**daption by **U**ncertainty estimation (TAU).

serious consequences such as gender bias (Savoldi et al., 2021). There are two kinds of uncertainty proposed by previous research: aleatoric uncertainty and epistemic uncertainty (Der Kiureghian and Ditlevsen, 2009; Kendall and Gal, 2017). While aleatoric uncertainty pertains to data noise in observations and cannot be easily eliminated, epistemic uncertainty stems from the insufficient knowledge of a model. Given that the training data for neural metrics primarily revolves around the News domain, this paper focuses on reducing epistemic uncertainty, particularly for out-of-distribution data.

Test-time Adaptation Domain adaptation (Pan and Yang, 2010) offers a deterministic target and can be trained with additional data through supervised or unsupervised methods, providing an intuitive approach to reduce epistemic uncertainty. However, there is a dearth of research exploring domain adaptation in MT evaluation due to the scarcity of multi-domain human ratings. Another limitation of using domain adaptation methods to mitigate epistemic uncertainty is the unknown domain of input data in real-world scenarios. This becomes particularly crucial for neural metrics, as they need to score diverse inputs without introducing domain bias. Test-time adaptation paradigm handles this challenge as a viable solution and can be categorized into test-time training (Sun et al., 2020) and source-free test-time adaptation (Kundu et al., 2020; Liang et al., 2020; Wang et al., 2021). It generalizes the model to out-of-distribution data during the testing phase without necessitating additional fine-tuning operations. Notably, a concurrent work (Lee and Lee, 2023) in the image classification tasks has also proposed minimizing uncertainty during test time. However, there are notable distinctions between our approach and theirs in terms of learning objectives and the specific type of uncertainty being targeted. In the context of MT evaluation, we present the first application of this paradigm and contribute a novel method that minimizes epistemic uncertainty at test time.

3 Method

The proposed method, as illustrated in Figure 2, is comprised of three distinct stages. These stages will be thoroughly discussed in the following section. Since both reference-based regression model and quality estimation (QE) model are used in COMET framework, we use $\langle h, s, \cdot \rangle$ to denote input data in order to take two major types of metrics mentioned in Section 2.

3.1 Uncertainty Estimation

The uncertainty is widely used in the classification model to obtain confidence about the classification results over a distribution P. Due to the fact that most neural metrics are regression models instead of classification model, for an input $\langle h, s, \cdot \rangle$, the regression model only produce a single score q rather than a score distribution $P(\mathbf{q})$. Therefore, it is a non-trivial question that how can obtain score distribution P. Glushkova et al. (2021) highlighted that Monte Carlo Dropout (MCD; Gal and Ghahramani, 2016) and Deep Ensemble (DE; Lakshminarayanan et al., 2017) are two approaches

used for estimating the uncertainty of a regression model. DE involves using multiple models that vary in randomization methods to predict scores for the same input, and then aggregating them to obtain a scoring distribution. Similarly, MCD also relies on models with different randomization, but only requires a single model with dropout enabled (Srivastava et al., 2014). The dropout technique introduces randomness by altering the activation status of model parameters during inference, simulating the effects of multiple homologous models used in DE.

Since our method focuses on adapting a single model to the target distribution, we choose MCD to estimate the score distribution due to its convenience and relatively low computational cost. Specifically, given an input $\langle h, s, \cdot \rangle$ and a model parameterized with θ, MCD makes model perform K-times feed-forward pass with different sets of parameters θ_k to get a score distribution $P(\mathbf{q}) = \{\mathrm{M}(\langle h, s, \cdot \rangle; \theta_k)\}_{k=1}^{K}$. Subsequently, the uncertainty can be calculated by the variance of score distribution $P(\mathbf{q})$, which can be formally expressed as:

$$u(\langle h, s, \cdot \rangle) = \mathbf{Var}(\{\mathrm{M}(\langle h, s, \cdot \rangle; \theta_k)\}_{k=1}^{K}) \quad (3)$$

where $\mathbf{Var}$ is the calculation process of variance. We use the standard deviation in implementation:

$$\mathbf{Var}(P) = \sqrt{\mathbb{E}\left[(P - \mu_P)^2\right]} \quad (4)$$

3.2 Adaptation by Uncertainty Minimization

After acquiring the model uncertainty through the methodologies outlined in the preceding section, it is advisable to expand the estimation procedure from instance-level to batch-level and run the estimation method in parallel. This approach serves two purposes: firstly, it enables seamless integration of the proposed method with the original inference process; secondly, it promotes stability in the optimization process by incorporating batch-level characteristics. Utilizing the uncertainty of each sentence independently as a guide for optimizing the model parameters would hinder the acquisition of adequate domain-specific features and potentially lead to a compromised starting point. To circumvent these challenges, the adaptation algorithm is designed at the batch level.

Another crucial problem is the choice of optimization parameters. Despite the existing categorization of data instances into different domains within the benchmark, there still exist differences among these domain-specific instances (Moore and Lewis, 2010). To deal with this problem, we ought to make the optimization process flexible to switch between different batches but not deviate too far from the original representation. Therefore, we choose to optimize a small fraction of the original model parameters, including the layer-wise attention and the corresponding coefficients.

The architecture of neural metric model typically consists of a pre-trained encoder and a score estimator, as illustrated in Figure 2. The score estimator is responsible for regression-based prediction of score q and takes the sentence embedding $\mathbf{O}_{\text{embed}}$ generated by L-layer[1] encoder as its input. In the COMET framework, the sentence embedding is obtained by aggregating the output $\mathbf{h}_i$ of each layer using layer-wise attention $\mathbf{w} = \{w_i\}_{i=1}^{L}$, which can be formulated as follows:

$$\mathbf{O}_{\text{embed}} = \gamma \cdot \sum_{l=1}^{L} w_i \cdot \mathrm{LayerNorm}(\mathbf{h}_i) \quad (5)$$

where γ is a learnable scaling coefficient and $\mathrm{LayerNorm}(\cdot)$ denotes layer normalization operation (Ba et al., 2016). Therefore, it is intuitive to achieve flexible adaptation by influencing the computation of sentence embedding, given its pivotal role in comprehending the semantic aspects of the text. We choose γ and $\mathbf{w}$ as the optimization parameters θ^*. For the empirical exploration of other optimization choices, we leave the discussion of this question in Section 5.1.

Algorithm 1 outlines the process of test-time adaptation by uncertainty minimization (TAU) when evaluating a specific MT system. The batch-level optimization, as described in the fifth to the eighth line, aligns with the aforementioned explanations. However, a notable challenge arises during the initial stages of optimization, commonly known as the "cold start" problem, if the test set is traversed only once. At the beginning of optimization, the model estimates the uncertainty using a small portion of the data, which prevents the early samples from benefiting from test-time adaptation compared to subsequently encountered samples. Therefore, the proposed method considers performing multiple adaptations for the entire system-level data, as indicated in the third line of Algorithm 1. In this way, the well-adapted model can re-score

[1]For the XLM-R model used by COMET framework, L is set to 24.

Algorithm 1 TAU: Test-time Adaptation by Uncertainty Minimization

Require: Model θ, System-level evaluation tuple $\mathcal{D} = \{\langle \mathbf{h}, \mathbf{s}, \cdot \rangle\}$, Adaptation rate α, Adaptation times J.

1: Backup original model $\theta^{'} \leftarrow \theta$
2: Select parameters for adaptation $|\theta^*| \ll |\theta|$
3: **for** adaptation iteration $j = 1, ..., J$ **do**
4: Score set $\mathbf{q} = \{\emptyset\}$
5: **for** mini-batch $\{\langle h, s, \cdot \rangle\}_{i=1}^{N} \in \mathcal{D}$ **do**
6: Estimate uncertainty u by Equation 3
7: Optimize $\theta^* \leftarrow \theta^* - \alpha \nabla_{\theta^*} \frac{1}{N} \sum_{i=1}^{N} u_i$
8: **end for**
9: Infer score $[q]$ by Equation 7
10: $\mathbf{q} \Leftarrow [q]$
11: **end for**
12: Restore to original model $\theta \leftarrow \theta^{'}$
13: **return** $\mathbf{q}$

the previous samples that may receive an uncertain score suffered by the cold start problem.

To conclude, the optimization objective of TAU can be formally expressed as follows:

$$\theta^* = \arg\min_{\theta^*} \mathbb{E}_{\langle h,s,\cdot \rangle \in \mathcal{D}} \left[u(\langle h, s, \cdot \rangle) \right] \quad (6)$$

3.3 Inference

Although the mean of the score distribution $P(\mathbf{q})$ estimated by MCD process can be viewed as a prediction score, it is not adopted in order to ensure comparability with other baseline models. Consequently, the inference stage of the adapted model aligns with conventional inference practices. To achieve this, the adapted model does not employ back-propagation of gradients and dropout during the inference process, as stated in the 9th line of Algorithm 1. The inference process can be formulated as follows:

$$q = \mathrm{M}_{\theta + \Delta\theta^*}(\{\langle h, s, \cdot \rangle\}) \quad (7)$$

In summary, the model leverages the MCD to estimate prediction uncertainty u of current data $\mathcal{D}$. This uncertainty serves as a signal to update the partial parameters θ^* during test time, ultimately leading to self-corrected predictions. Moreover, the update process is performed online, ensuring that no additional storage costs are incurred.

4 Experiments

4.1 Experimental Setups

Data We conduct experiments on a multi-domain benchmark of WMT21 Metrics Task[2], which includes three language pairs and corresponding MQM scores. Compared to previous WMT crowd-sourced evaluations, MQM framework is a more granular evaluation protocol that focuses on explicit errors. Freitag et al. (2021a) explored the application of the MQM framework (Lommel et al., 2014) in the evaluation of WMT submissions and published an alternative set of reference scores annotated by human experts[3]. We used MQM scores as the reference and evaluate how well the scores produced by metrics correlate with them. For News domain that has multiple references, we extend the evaluation of metrics to include human translations (HT) alongside the standard reference. It is important to note that HT is out-of-distribution data for neural metrics, given that these metrics have primarily been trained on the scoring data related to existing MT systems. Specifically, the metrics need to conduct the system-level evaluation by involving (*w/ HT*) or excluding HT text (*w/ HT*).

Baselines The baselines cover three mainstream types of metrics:

- **Text-based Metrics**: Traditional metrics quantify the n-gram overlap between the hypothesis and reference, such as BLEU (Papineni et al., 2002) and CHRF (Popović, 2015), or measure the edit distance like TER (Snover et al., 2006). These metrics employ transparent evaluation mechanisms that draw inspiration from human evaluation. However, their scope is limited to assessing the surface-level coverage at the morphological level.

- **Embedding-based Metrics**: The evaluation process of embedding-based metrics is also transparent and characterized by strong interpretability. These metrics measure the semantic-level similarity between reference and hypothesis embeddings, which are encoded using a pre-trained encoder or language model (Devlin et al., 2019). This approach provides a more nuanced evaluation perspective compared to text-based metrics. Among

[2]https://www.statmt.org/wmt21/metrics-task.html
[3]https://github.com/google/wmt-mqm-human-evaluation/

Metrics	News w/o HT			News w/ HT			TED			Avg.
	En-De	Zh-En	En-Ru	En-De	Zh-En	En-Ru	En-De	Zh-En	En-Ru	
Baselines										
TER	93.0	41.6	-4.1	7.4	-8.5	-28.9	50.6	42.1	69.7	29.2
BLEU	93.7	31.0	50.7	13.2	-15.2	-4.3	62.0	32.4	82.8	38.5
CHRF	89.8	30.2	78.3	1.7	-14.3	12.3	47.1	36.3	82.5	40.4
BERTSCORE	93.0	54.2	62.9	7.4	9.5	-12.3	50.6	30.6	83.1	42.1
COMET-DA$_{2020}$	81.4	51.1	67.6	65.8	22.1	55.6	78.8	25.1	85.9	59.3
COMET-MQM-QE$_{2021}$	71.1	52.9	63.2	79.2	61.9	68.1	69.4	-20.9	88.4	59.3
COMET-MQM$_{2021}$	77.1	62.8	65.9	72.0	33.6	68.5	81.8	26.6	84.1	63.6
Reproduced Results and Our Methods										
♢ COMET-DA$_{2020}$	81.5	51.1	67.5	58.0	26.4	56.8	78.8	25.0	85.9	59.0
+TAU	**85.7**	**53.5**	**71.0**	48.0	**27.4**	54.5	**85.9**	**28.3**	**87.3**	**60.2**
♢ COMET-MQM-QE$_{2021}$	71.2	53.0	68.8	79.2	61.9	68.1	69.4	-20.8	81.7	59.2
+TAU	62.8	**57.4**	**70.3**	72.0	**65.2**	**78.1**	**82.9**	**25.7**	80.7	**66.1**
♢ COMET-MQM$_{2021}$	77.2	62.8	65.9	69.8	48.7	69.7	81.8	26.6	84.1	65.2
+TAU	76.5	**69.2**	**67.1**	**75.4**	**67.8**	**71.4**	**87.5**	24.5	**84.9**	**69.4**

Table 1: System-level Pearson correlations of metrics with MQM scores on available language pairs for WMT21 Metrics Shared Task. **Bold** values indicate that the model receives improvement by applying our proposed method. "Avg." denotes averaged results.

them, the representative BERTSCORE (Zhang et al., 2020) metric is used in our experiments.

- **Neural Metrics**: Since the evaluation mechanism of neural metric has been described in Section 2, we will not go into details in this part. There are several models provided in COMET framework (Rei et al., 2020, 2021) including reference-based and reference-free models. We choose three representative models as the baselines and testbed: COMET-DA$_{2020}$, COMET-MQM$_{2021}$ and COMET-MQM-QE$_{2021}$, where the last one only requires source text to evaluate the translation.

The reported performance of baselines is taken from official results (Freitag et al., 2021b). To minimize the possible bias in our experiments, we reproduced COMET baselines using open-sourced repository[4] and implement our method on the same code skeleton.

Settings During the process of test-time adaptation, the learning rate α is set to $1e-4$ by using WMT20 benchmark as the development set. We only tune the batch size N and adaptation times J for better performance. We use Adam optimizer (Kingma and Ba, 2015) to update parameters θ^* with $\beta_1 = 0.9, \beta_2 = 0.99$ and $\epsilon = 10^{-8}$. For estimating the uncertainty, we perform feed-forward operation $K = 30$ times with dropout enabled.

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

4.2 Meta-Evaluation

To assess the system-level performance of the metric, we employ two meta-evaluation methods: correlation performance and pairwise accuracy. The Pearson correlation, renowned for its widespread application, serves as a common metric used in evaluating system-level performance. This measurement has also been adopted by the WMT Shared Task as a means to evaluate the performance of metrics. In addition, pairwise accuracy (Kocmi et al., 2021) measures how many system pairs are correctly ranked by the metric, which can be calculated as follows:

$$\text{Acc.} = \frac{|\text{sign}(\text{metric}\Delta) = \text{sign}(\text{human}\Delta)|}{|\text{system pairs}|} \quad (8)$$

where Δ and $\text{sign}(\cdot)$ denote the differences and the sign function, respectively. While most existing work calculates the correlation (e.g., Pearson correlation) between metric scores and human judgments to evaluate their performance, a reliable metric should also be able to correctly compare and rank MT systems. Therefore, we report pairwise accuracy in addition to Pearson correlation performance to demonstrate the system-level ranking performance, serving as a cross-validation metric.

Metrics	News w/o HT			News w/ HT			TED			Avg.
	En-De	Zh-En	En-Ru	En-De	Zh-En	En-Ru	En-De	Zh-En	En-Ru	
♢ COMET-DA$_{2020}$	82.1	70.5	68.1	72.4	61.5	66.7	82.1	69.2	82.4	72.8
+TAU	**89.7**	69.2	**73.6**	**76.2**	59.3	**70.5**	**85.9**	67.9	**83.5**	**75.1**
♢ COMET-MQM-QE$_{2021}$	73.1	78.2	69.2	78.1	81.3	73.3	71.8	41.0	80.2	71.8
+TAU	71.8	75.6	**75.8**	77.1	79.1	**79.0**	**80.8**	**57.7**	80.2	**75.3**
♢ COMET-MQM$_{2021}$	79.5	66.7	68.1	77.1	61.5	70.5	87.2	66.7	78.0	72.8
+TAU	**83.3**	66.7	64.8	**80.0**	**63.7**	68.5	**88.5**	65.4	**82.4**	**73.7**

Table 2: Pairwise accuracy of metrics with MQM ranking results on available language pairs for WMT21 Metrics Shared Task. **Bold** values indicate that the model receives improvement or maintains the performance by applying our proposed method. Among them, identical results are underlined. "Avg." denotes averaged results.

We use functions from `mt-metrics-eval`[5] toolkit to calculate the above two meta-evaluation results.

4.3 Main Results

As can be seen from Table 1, the proposed method TAU partially improves the averaged correlation performance of COMET metrics, and the improvements vary from model to model. Models trained on MQM scores demonstrate a greater benefit from adaptation compared to COMET-DA models whose training data is direct assessment (DA) scores. This observation suggests that TAU exhibits characteristics akin to continual learning when the test data is related to the training data source. The cross-validation results in Table 2 show a similar tendency as what is observed in Table 1. Since we did not perform hyper-parameter searching on pairwise accuracy, which further supports the effectiveness of the proposed method. From a model-level comparison standpoint, the QE model still receives larger improvements. However, it is notable that adaptation may occasionally result in a performance decline. Therefore, the decision to do adaptation or not becomes a vital consideration for in-domain data, and the subsequent section will delve into the effect of distribution differences through an empirical study.

5 Analysis

In this section, we will discuss the effectiveness of our method by answering three questions: 1) How do different optimization settings impact performance? 2) When does test-time adaptation work? 3) Can the proposed method effectively reduce epistemic uncertainty? Among these questions, the last one serves to justify our research objective and entails a segment-level analysis to understand why the proposed method is effective.

[5]https://github.com/google-research/mt-metrics-eval/

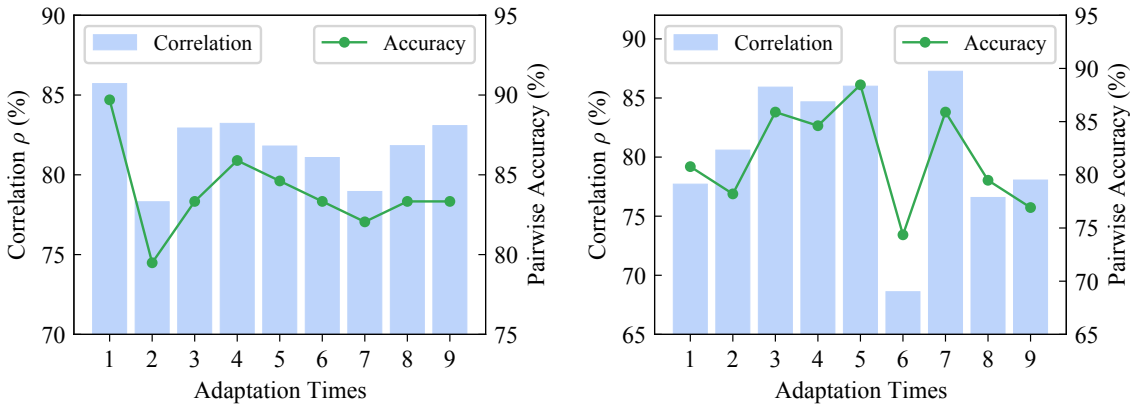
Domain	LAtt.	LN.	Estim.	ρ	Acc.
News	✓	✗	✗	**85.7**	**89.7**
	✗	✓	✗	79.5	76.9
	✗	✗	✓	78.7	80.8
	✓	✓	✗	79.6	76.9
	✓	✗	✓	78.6	80.8
	✓	✓	✓	78.0	79.4
TED	✓	✗	✗	**85.9**	**85.9**
	✗	✓	✗	79.4	82.1
	✗	✗	✓	77.2	76.9
	✓	✓	✗	79.4	82.1
	✓	✗	✓	77.1	76.9
	✓	✓	✓	76.9	76.9

Table 3: Performance of TAU when optimizing different modules. "LAtt.", "LN." and "Estim." denotes layerwise attention, layer normalization, and estimator. Overall, optimizing layerwise attention is a suitable choice.

5.1 Ablation Study

We use COMET-DA model to conduct an ablation study since it was not tuned for MQM scoring.

Adaptation Parameters Table 3 reveals that the parameters of layerwise attention module are suitable to optimize at test time, addressing the concerns raised in Section 3.2. The conducted comparisons reveal that optimizing parameters other than the layerwise attention module ultimately results in performance degradation. This degradation persists even when jointly tuning with the layerwise attention module. These findings confirm our initial hypothesis that optimization should not deviate too far from the original parameters, thereby avoiding extensive optimization of core components or a larger number of parameters. A closer examination of the degree of performance degradation indicates that optimizing the estimator produces the most



(a) News (b) TED

Figure 3: Performance of TAU with different settings of adaptation times. The out-of-distribution data requires more adaptation times than in-domain data, and both of them would suffer from extreme settings.

significant decline in performance, aligning with the aforementioned reasons.

Adaptation Times To address the "cold start" problem discussed earlier, Algorithm 1 incorporates a multiple adaptation policy. The empirical results presented in Figure 3 reveal a relationship between the choice of adaptation times and the domain. Specifically, in-domain data (News) suffers from continuous adaptation, whereas out-of-distribution data (TED) demonstrates improved performance through multiple adaptations. In the case of in-domain data, the data shift between training and inference is relatively smaller compared to out-of-distribution data, allowing the performance to reach its peak with fewer adaptation runs. In contrast to in-domain behaviors, optimizing out-of-distribution data takes longer due to the need for dissimilar data features, leading to fluctuations in performance indicators. Nevertheless, a common trend emerges where larger adaptation times eventually hinder performance, particularly for in-domain data. To strike a balance between computational time and performance, all the adaptation times utilized in the previous experiments are limited to no more than 5 times.

5.2 Effects of Data Types

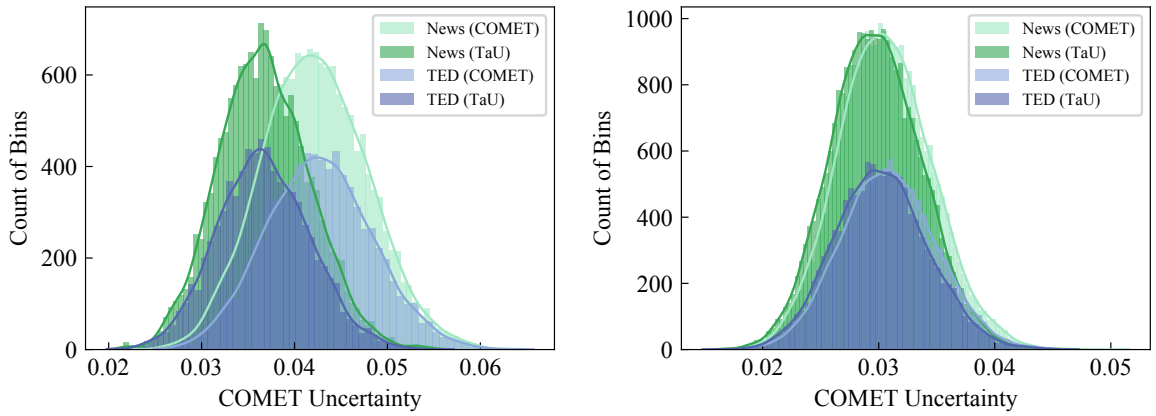
In order to determine which type of data benefits more from the proposed method TAU, we categorize the evaluation tasks into three distinct types, and then report the performance changes for each type in Table 4. The scope of out-of-distribution data extends beyond TED data from out-of-domain sources, encompassing human translations (HT) as well. The human translations rarely present in training data and differ significantly from the text generated by MT systems. Thus, the tasks within "News w/HT" category are regarded as partial out-of-distribution scenarios. Overall, the proposed method achieves the highest improvement for each model when evaluated on out-of-distribution data, as evidenced by the average correlation metric. It is plausible because a major source of uncertainty is out-of-distribution data, and TAU is able to alleviate inference bias in these cases.

Models	ΔID.	ΔPartial OD.	ΔOD.
DA	3.4	-3.8	**4.0**
MQM	2.4	**8.8**	1.5
QE-MQM	-0.8	2.0	**19.7**
Avg.	1.7	2.4	**8.4**

Table 4: Performance variation of different data types for each model. Δ, "ID.", "OD." represent performance changes, in-domain, out-of-distribution, respectively.

5.3 Model Uncertainty

In response to our research objectives, we investigate whether the uncertainty has been reduced after applying the proposed method. We aggregated the uncertainty values at the segment level and visualized their distributions grouped by languages and models, as depicted in Figure 4. These visualizations demonstrate a shift in the distributions for both in-domain and out-of-distribution data, affirming the effectiveness of uncertainty minimization. However, it is worth delving into the reasons behind the larger uncertainty shift observed in



(a) English-German (COMET-DA$_{2020}$)

(b) Chinese-English (COMET-MQM$_{2021}$)

Figure 4: Uncertainty distribution of COMET baselines and corresponding models optimized by TAU. TAU can reduce the uncertainty of different models on different domains.

COMET-DA model compared to COMET-MQM model. The discrepancy could be attributed to the training data. COMET-MQM model is derived by fine-tuning COMET-DA model on normalized MQM scores, which employ a scoring protocol that deviates from the traditional point-wise scale. Specifically, the segment-level MQM score is derived from the count of explicit errors and ranges from -25 to 0, unlike the continuous $[1, 100]$ scale adopted by WMT (Freitag et al., 2021a). We observed that there are many identical scores such as "0", which means that the annotators consider them to be perfect translations. As a consequence, the MQM scores exhibit less diversity compared to the DA scores, subsequently influencing the prediction behavior of the models fine-tuned on MQM scores. Encouragingly, despite these factors, we were able to reduce uncertainty in the MQM models and improve their overall performance.

6 Conclusion

The uncertainty of neural metrics is proven to be associated with prediction error and limits generalizing them for a wider range of applications. In this paper, we propose a novel method, TAU, to minimize the uncertainty of neural metrics at test time in unsupervised settings without learning extra data. Our experimental results showcase the efficacy of TAU in reducing test-time uncertainty while simultaneously improving the performance of widely used metrics. In addition, our findings indicate that the proposed method exhibits enhanced effectiveness when applied to out-of-distribution data in comparison to in-domain data, which lays a solid foundation for its potential application to other models. However, the segment-level performance does not significantly outperform the baselines. In the future, we will polish the methods for better segment-level correlation performance and explore the test-time adaptation on large language models across various tasks.

Limitations

The methodology and experimental approach presented in this paper have certain limitations concerning their practical application and the availability of language resources. The proposed method estimates uncertainty using Monte Carlo Dropout with K iterations and subsequently performing adaptation J times. These additional computations result in increased inference time in real-world applications. Empirical evidence suggests that larger values of J lead to a linear increase in time costs in practical scenarios. Although the number of J on the WMT21 benchmark was limited in our experiments, the exact cost associated with achieving successful adaptation for new models or datasets remains uncertain. In terms of language resources, the majority of MT metric benchmarks still focus on the News domain, leaving a dearth of multi-domain MQM benchmarks for conducting more meta-evaluation experiments during the preparation of this paper. To address these limitations, it is imperative to explore the performance of the proposed methodology on a wider range of out-of-distribution benchmarks in the future. Furthermore, as highlighted by the reviewer, it is also important to note that the proposed methodology does not

consistently exhibit performance improvements on certain specific test sets. One possible explanation for this observation could be attributed to our investigation of the optimal learning rate using the WMT20 dataset. The divergence in scoring perspectives between the conventional WMT score and the MQM score might lead to discrepancies in improvement trends.

Ethics Statement

An ethical concern associated with neural metrics is the presence of unpredictable bias in the evaluation process. Unlike traditional text-based metrics, neural metrics pose challenges in mitigating evaluation bias due to their black-box nature, which also introduces potential issues like gender bias inherent in pre-trained language models. While our current study does not investigate the bias problem, reducing uncertainty in the evaluation process may help contribute to mitigating the potential risks associated with generating biased results.

Acknowledgment

This work was supported in part by the Science and Technology Development Fund, Macau SAR (Grant Nos. FDCT/060/2022/AFJ, FDCT/0070/2022/AMJ), the National Natural Science Foundation of China (Grant No. 62206076), the Research Program of Guangdong Province (Grant No. 2220004002576), Shenzhen College Stability Support Plan (Grant Nos. GXWD20220811173340003, GXWD20220817123150002), Shenzhen Science and Technology Program (Grant No. RCBS20221008093121053) and the Multi-year Research Grant from the University of Macau (Grant No. MYRG2020-00054-FST). This work was performed in part at SICC which is supported by SKL-IOTSC, and HPCC supported by ICTO of the University of Macau. We would like to thank the anonymous reviewers and meta-reviewer for their insightful suggestions.

References

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *ArXiv preprint*, abs/1607.06450.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112.

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.

Erick Fonseca, Lisa Yankovskaya, André F. T. Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 1–10, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021a. 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, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021b. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André F. T. Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5574–5584.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

Jogendra Nath Kundu, Naveen Venkat, Rahul M. V., and R. Venkatesh Babu. 2020. Universal source-free domain adaptation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 4543–4552. IEEE.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6402–6413.

JoonHo Lee and Gyemin Lee. 2023. Feature alignment by uncertainty and self-training for source-free unsupervised domain adaptation. *Neural Networks*, 161:682–692.

Jian Liang, Dapeng Hu, and Jiashi Feng. 2020. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6028–6039. PMLR.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Revista Tradumàtica: tecnologies de la traducció*, 12:455–463.

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.

Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online. Association for Computational Linguistics.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. 2018. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 61–72, Brussels, Belgium. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

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, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. 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.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

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.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9229–9248. PMLR.

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.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. 2021. Tent: Fully test-time adaptation by entropy minimization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, and Lidia S. Chao. 2021a. Difficulty-aware machine translation evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 26–32, Online. Association for Computational Linguistics.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, and Lidia S. Chao. 2021b. Variance-aware machine translation test sets. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Discussed in the "Limitation" section after the conclusion but before the references.

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract, Section 1: Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Not applicable. Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☑ Did you run computational experiments?

Section 4, 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4.1

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4.3, 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4.1

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.