Partial Could Be Better Than Whole. HW-TSC 2022 Submission for the Metrics Shared Task

Yilun Liu*, **Xiaosong Qiao***, **Zhanglin Wu, Chang Su, Min Zhang,**
Yanqing Zhao, Song Peng, Shimin Tao, Hao Yang, Ying Qin,
Jiaxin Guo, Minghan Wang, Yinglu Li, Peng Li, Xiaofeng Zhao
Huawei Translation Services Center, Beijing, China
{liuyilun3, qiaoxiaosong, wuzhanglin2, suchang8, zhangmin186, zhaoyanqing, pengsong2, taoshimin, yanghao30, qinying, guojiaxin, wangminghan} @huawei.com

Abstract

In this paper, we present the contribution of HW-TSC to WMT 2022 Metrics Shared Task. We propose one reference-based metric, HWTSC-EE-BERTScore*, and four reference-free metrics including HWTSC-Teacher-Sim, HWTSC-TLM, KG-BERTScore and CROSS-QE. Among these metrics, HWTSC-Teacher-Sim and CROSS-QE are supervised, whereas HWTSC-EE-BERTScore*, HWTSC-TLM and KG-BERTScore are unsupervised. We use these metrics in the segment-level and system-level tracks. Overall, our systems achieve strong results for all language pairs on previous test sets and a new state-of-the-art in many sys-level case sets.

1 Introduction

Due to the expensive cost of manual evaluation, automatically evaluating the outputs of translation systems is critically important in the field of machine translation (MT) (Freitag et al., 2021a). Therefore, a lot of automatic metrics have been proposed to approach this task. According to whether the reference sentences are required or not, the metrics are categorized into two classes: (1) reference-based metrics like BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), BERTScore (Zhang et al., 2020) and BLEURT (Sellam et al., 2020), which evaluate the hypothesis by referring to the golden reference; (2) reference-free metrics like YiSi-2 (Lo, 2019) and COMET-QE (Rei et al., 2020, 2021), which are also referred as quality estimation (QE). These metrics estimate the quality of hypothesis only based one source sentences without using references.

In this paper, we present the contribution of HW-TSC to the WMT 2022 Shared Task on Metrics. We participated in the segment-level and system-level tracks with 1 reference-based metric (HWTSC-EE-BERTScore*) and 4 reference-free metrics (HWTSC-Teacher-Sim, HWTSC-TLM, KG-BERTScore and CROSS-QE). Details of our metrics are illustrated in Table 1.

HWTSC-EE-BERTScore* (Entropy Enhanced Metrics)is built upon existing metrics, aiming to achieve a more balanced system-level rating by assigning weights to segment-level scores produced by backbone metrics. The weights are determined by the difficulty of a segment, which is related to the entropy of a hypothesis-reference pair. A translation hypothesis with a significantly high entropy value is considered difficult and receives a large weight in aggregation of EE-Metrics' system-level scores.

HWTSC-Teacher-Sim is a supervised reference-free metric with the framework of BERTScore (Zhang et al., 2020), which is obtained by fine-turning the multilingual Sentence-BERT model (Reimers and Gurevych, 2019, 2020a). Both the unsupervised TearcherSim (Yang et al., 2022b,a) and the implicit multilingual word embedding alignment (Zhang et al., 2022b) have shown that the pretained multilingual Sentence-BERT model is very effective for both reference-based and reference-free MT evaluations on WMT DA (Direct Assessment) data. However, its performance on WMT MQM (Multidimensional Quality Metrics) data is poor. We propose an effective training strategy for the pretrained multilingual Sentence-BERT and a novel normalization method for the DA and MQM scores.

HWTSC-TLM (Zhang et al., 2022a) is an unsupervised reference-free metric which only uses the system translations as input and calculates the scores by a target-side language model. Although source sentences are not considered, the results of this metric with XLM-R (Conneau et al., 2020) on WMT19 are very promising.

KG-BERTScore (Wu et al., 2022) is an unsupervised reference-free metric, which incorporates multilingual knowledge graph into BERTScore

* equal contribution

Proceedings of the Seventh Conference on Machine Translation (WMT), pages 549 - 557
December 7-8, 2022

Metrics	Reference	Training	Segment-level	System-level
HWTSC-EE-BERTScore*	reference-based	unsupervised	✗	✓
HWTSC-Teacher-Sim	reference-free	supervised	✓	✓
HWTSC-TLM	reference-free	unsupervised	✓	✓
KG-BERTScore	reference-free	unsupervised	✓	✓
CROSS-QE	reference-free	supervised	✓	✓

Table 1: Description of 5 metrics participated in WMT 2022 Shared Task. ✓and ✗respectively indicate whether the metric participates the corresponding track or not.

(Zhang et al., 2020). The score of this metric is calculated by linearly combining the results of BERTScore and bilingual named entity matching.

CROSS-QE is an application of "QE as a metric". Based on our previous work (Yang et al., 2020; Wang et al., 2020; Chen et al., 2021), we propose a reference-free metric, like COMET-QE architecture.

2 Metrics

This section introduces our metrics for WMT Metrics 2022 Shared Task including Reference-based and reference-free.

2.1 Reference-based

This year, entropy-enhanced BERTScore (HWTSC-EE-BERTScore, or referred as EE-BERTScore in short) was used in the general tests of the system-level track. EE-BERTScore, built upon standard BERTScore (Zhang et al., 2019), is within one of the EE metrics proposed earlier (Liu et al., 2022). The main idea of EE metrics is to challenge the standard way of acquiring system-level scores that outputs a simple arithmetic average of scores on segments in the evaluation set, and to provide a framework that enhances existing MT metrics by assigning higher weights to the difficult samples in the evaluation set. The motivation is simple: for MT evaluation, it is not likely that human raters treat every source-reference pair equally. Those simple samples can be easily translated, leading to similar human scores given to different hypotheses, while the more challenging part in an evaluation set often distinguishes top candidates from inferior systems. Like different weights are assigned to questions in real-world examinations based on variant difficulties, MT evaluation metrics should also encourage systems that perform better on relatively difficult samples. In the preliminary experiment, we find that using only the difficult segments (usually counting for less than 5% of all segments in the whole evaluation set) to evaluate MT systems, doesn't lead the automatic metrics to give incorrect ratings for MT systems, and sometimes even improves the performances of metrics in terms of correlation with human DA scores. Thus, we proposed EE metrics, which emphasize the translation qualities of relatively difficult ones among all hypotheses given by a system and assign high weights to these hypotheses in the aggregation of system-level scores.

2.1.1 Working Process of EE Metrics

Currently, EE metrics determine the difficulty of a segment via the average qualities of hypotheses. The qualities are measured by the translation entropy (or chunk entropy) (Yu et al., 2015) between the reference and the hypothesis. For a human reference and a hypothesis given by an MT system, a high chunk entropy suggests high uncertainty of the translation (the more linguistically matched parts between the hypothesis and the reference is, the lower the uncertainty of the translation is) and a low entropy indicates good confidence of the given hypothesis in expressing the meaning of the source segment. For example, if a hypothesis is perfectly matched with a reference, then the entropy of the translation is zero, and if there is no matching token between the hypothesis and the reference, the chunk entropy is positive infinity, indicating a total uncertainty and disorderness of the translation.

Fig. 1 illustrates how EE metrics assign different weights to the segments in the evaluation set based on the computed entropy. Firstly, segments in the evaluation set are divided into two groups: easy samples and difficult samples. If the entropy of a hypothesis is higher than the threshold h, it is considered in the difficult group and vice versa. Then, hypotheses are assigned weights in the aggregation of final score based on the groups they belong to. Specifically, samples in the easy group receive a weight of w/N_e and samples in the difficult group

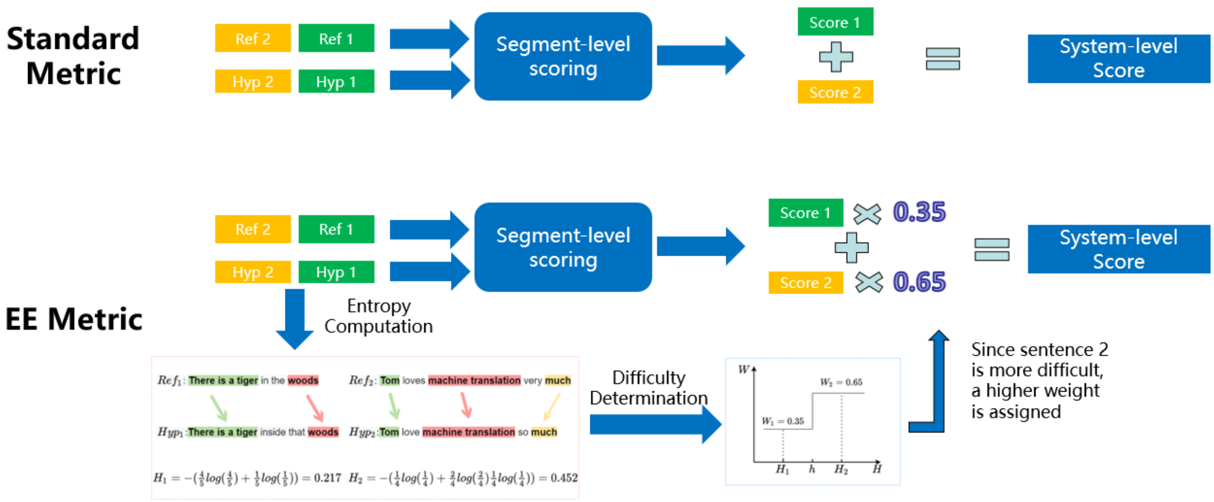

Figure 1: Workflow of EE metrics, assuming the evaluation set contains two segments with reference-hypothesis pairs (Hyp 1, Ref 1) and (Hyp 2, Ref 2).

receive a weight of $(1-w)/N_d$, where N_e, N_d are the sizes of easy and difficult group, respectively, and w is a balance coefficient that, in our earlier version of EE metrics, may vary for different language pairs and evaluation datasets. Since the number of easy hypothesis is much larger than the number of difficult hypothesis for a given MT system, the weight of easy samples is much lower than the weight of difficult samples.

2.1.2 EE Metrics 2.0 vs. EE Metrics 1.0

The earlier version of EE metrics (denoted as EE metrics 1.0) has two hyper-parameters: h and w, involving in the selection of difficult samples and the determination of weights assigned to each group, respectively. The existence of such hyper-parameters hinders the application of EE metrics. What's worse, the hyper-parameters often alter for different language pairs and evaluation datasets (*e.g.*, we use up to 10 different parameters in our preliminary experiment, involving WMT 19 evaluation set), making it hard to estimate a feasible combination of parameters in the actual scenario. To alleviate such undesirable pain, we propose EE metrics 2.0 for this year's WMT metrics shared tasks. EE metrics 2.0 aims to reduce the hyper-parameters involved in the computation of system-level score as much as possible and offers a lightweight approach of computing weights for each segment. Specifically, EE metrics 2.0 doesn't require specifying h anymore, but automatically estimates thresholds based on a normal distribution fitting of average translation qualities (the average entropy) over all segments, aiming to find the threshold value of entropy where a sample has a significantly higher entropy than those of other samples in the datasets. Moreover, the estimation of w is simplified to a single value, instead of a series of different values for different language pairs. EE metrics 1.0 provides a formula to estimate w for every language pair, which is acquired based on the fitting of WMT 19 results. In contrast, the value of w doesn't change across different language pairs in EE metrics 2.0. Our submissions in WMT 2022 Metrics Shared Task contain three different configurations of values of w: 0.3, 0.5 and 0.8, which stand for different degrees of balance of weights received between difficult groups and easy groups.

2.2 Reference-free

In this section, we would introduce the four reference-free metrics.

2.2.1 HWTSC-Teacher-Sim

HWTSC-Teacher-Sim proposed by (Zhang et al., 2022b), is a Reference-free metric used for machine translation evalation by achieving cross-lingual word embedding alignment throgh multilingual knowledge distillation (MKD) (Reimers and Gurevych, 2020b). The procedure of multilingual knowleage distillation is described in the Figure 2. The teacher model is monolingual SBERT (Reimers and Gurevych, 2019) which achieves state-of-the-art performance for various sentence embedding tasks, and the student model is a multilingual pretrained model like mBERT or XLM-R before distillation. After MKD, the similarity score of sentence pairs in MT evaluation on the language

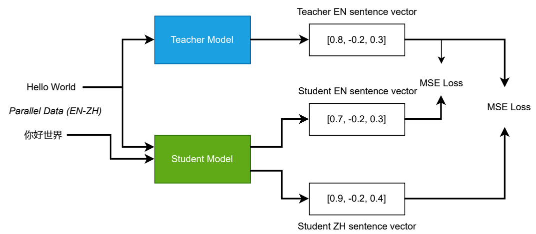

Figure 2: Multilingual knowledge distillation

model should be as high as possible. Based on this feature, embeddings of sentences are used to calculate the similarity score as a metric. And we achieve strong results using language models to calculate the similarity between sentence pairs in an supervised manner in MQM data.

2.2.2 HWTSC-TLM

HWTSC-TLM proposed by Zhang et al. (2022a) utilizes a pretrained multilingual model XLM-R (Conneau et al., 2020) to score the system translations, which is a zero-shot unsupervised metric for MT evaluation.

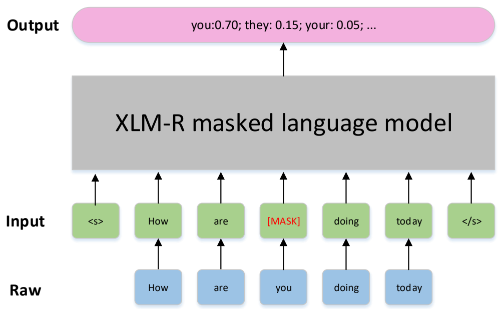

Figure 3: An example of HWTSC-TLM metric calculation for a given sentence

For a given sentence $\boldsymbol{s} = (w_1, \ldots, w_m)$ with m tokens, the score is defined as:

$$SEG\_LM(\boldsymbol{s}) = \frac{1}{m}\sum_{i=1}^{m} \log \frac{1}{P(w_i|\boldsymbol{s} - w_i)}, \quad (1)$$

where $P(w_i|\boldsymbol{s} - w_i)$ the probability of w_i predicted by the masked language model when w_i is replaced by `[MASK]`, as shown in Figure 3. And this score is used for segment-level MT evaluation.

For system-level evaluation where a set of system translation sentences S is provided, the score is defined as:

$$SYS\_LM(S) = \frac{1}{|S|}\sum_{\boldsymbol{s} \in S} SEG\_LM(\boldsymbol{s}), \quad (2)$$

which is the mean value of SEG_LM scores on each sentence in S.

2.2.3 CrossQE

CrossQE showed as figure 4 has used pre-trained Cross-lingual XLM-Roberta large(Lample and Conneau, 2019; Conneau et al., 2019) as predictor instead of RNN-based model in the two-stage Predictor-Estimator architecture (Kim et al., 2017), and uses regressor as quality estimator, and multitasks are trained at the same time. The Cross-lingual XLM-Roberta large model is pre-trained from large-scale parallel corpora which source and target tokens are concatenated by MLM task. Shuffling those tokens and predicting those tokens' index by the pre-trained model as a additional pre-training task can improve CrossQE's effect. CrossQE is build on the COMET architecture[1] by exploring adapter layers (Houlsby et al., 2019) for quality estimation to eliminate the over-fitting problem while instead of fine-tuning the whole base pre-trained model for different NLP tasks (He et al., 2021). At training step, the Mean Teacher loss(Baek et al., 2021) is added to improve model's over-fitting problem. Data augmentation method based on Monte Carlo (MC) dropout (Gal and Ghahramani, 2016) is added to enhance the performance in sentence quality score prediction.

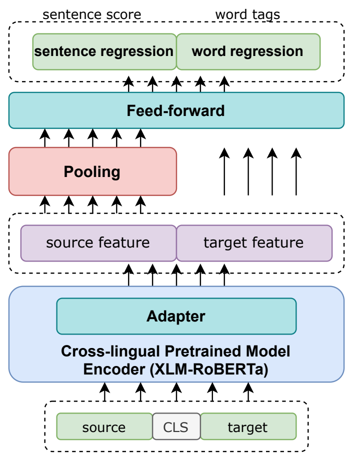

Figure 4: Cross QE architecture

2.2.4 KG-BERTScore

KG-BERTScore metric proposed by Wu et al. (2022), incorporates multilingual knowledge graph into BERTScore for reference-free MT evaluation. The evaluation process in WMT22 metrics shared task is shown in Algorithm 1:

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

Firstly, we employ a reference-free BERTScore metric to calculate F_{BERT} score of each MT sentence. For the WMT22 metrics shared task, we use HWTSC-Teacher-Sim metric to calculate F_{BERT} so that the score is more relevant to the MQM.

Secondly, we utilize model (NER) named entity recognition to identify named entities in the sentences, and retrieve the corresponding entity IDs in multilingual knowledge graph. We then calculate F_{KG} scores based on entity matching degree. Since the same named entities in different languages share the same entity ID in multilingual knowledge graph, we can check whether they can be matched by entity IDs. For the WMT22 metrics shared task, the NER model we use is spacy[2], and the multilingual knowledge graph we use is Google Knowledge Graph Search API[3].

Finally, we combine to obtain a segment-level $F_{KG-BERT}$ score, and the $F_{KG-BERT}$ score of all MT sentences are averaged to obtain a system-level score. For the WMT 2022 metrics shared task, we set α to 0.5, and if there is no entity in the source, F_{KG} score is 1.

In addition, due to limited access to the Google Knowledge Graph Search API, we only use KG-BERTScore metric to score the three language directions zh-en, en-ru, and en-de on the WMT22 metrics shared task. The scores for other language directions in our submissions are simply populated with the F_{BERT} score based on the paraphrase-multilingual-mpnet-base-v2 model[4].

Algorithm 1: KG-BERTScore evaluation process

Input : all source sentences $s_k \in S$ and machine translations $t_k \in T$ of n sentence pairs
Output : a system-level score F

```
for each sentence pair {s_k, t_k} ∈ {S,T} do
    // x_i, x_j, x̂_i, x̂_j is the word embedding.
    R_k = 1/|s_k| Σ_{x_i ∈ s_k} max_{x̂_j ∈ t_k} x_i^T x̂_j
    P_k = 1/|t_k| Σ_{x̂_i ∈ t_k} max_{x_j ∈ s_k} x̂_i^T x_j
    F_BERT_k = 2 (P_k · R_k) / (P_k + R_k)
    // entities(s_k), entities(t_k) is the number of entities.
    if entities(s_k) ≠ 0 then
        F_KG_k = matches(entities(s_k), entities(t_k)) / entities(s_k)
    else
        F_KG_k = 1
    end
    // α is an adjustable hyperparameter.
    F_KG-BERT_k = α · F_KG_k + (1 − α) · F_BERT_k
end
F = Σ_{k=1}^{n} F_KG-BERT_k / n
```

$$R_k = \frac{1}{|s_k|} \sum_{x_i \in s_k} \max_{\hat{x}_j \in t_k} x_i^T \hat{x}_j$$
$$P_k = \frac{1}{|t_k|} \sum_{\hat{x}_i \in t_k} \max_{x_j \in s_k} \hat{x}_i^T x_j$$
$$F_{BERT_k} = 2\frac{P_k \cdot R_k}{P_k + R_k}$$
$$F_{KG_k} = \frac{matches(entities(s_k), entities(t_k))}{entities(s_k)}$$
$$F_{KG-BERT_k} = \alpha \cdot F_{KG_k} + (1-\alpha) \cdot F_{BERT_k}$$
$$F = \frac{\sum_{k=1}^{n} F_{KG-BERT_k}}{n}$$

3 Experiments

3.1 Experiments of Reference-based

To verify the feasibility of EE metrics 2.0, we conduct experiments mainly on WMT 20 and WMT 21 using MQM (Lommel et al., 2014) as the ground truth. To investigate the difference between when human translations are used as a system and when they are not used, we display the results computed on two sets of systems for each language pair. We report three coefficients: Pearson's correlation r, Kendall's τ and Spearman's ρ, to validate system-level correlations with human evaluations.

Table 2 displays performance comparison between EE-BERTScore and standard BERTScore, where EE-BERTScore achieves overall higher correlations with human MQM than standard BERTScore. We experiment with EE-BERTScore under different values of w, suggesting different relative weights between easy groups and difficult groups in the computation of system-level scores. We find that each setting of w is able to improve the performance of standard BERTScore, and has their best performances on a certain dataset. For example, EE-BERTScore-0.3 and EE-BERTScore-0.5 achieve a strong result on news test of WMT 20 and WMT 21, while on WMT 21 tedtalks, best performance is achieved when w is 0.8.

Since EE metrics evaluate a system relying on not only the single system, but also other participated systems, the existence of human translations may have an impact on the performances of EE metrics. As shown in Table 2, correlations with MQM drop sharply for EE-BERTScore-$*$ when human translations are included as, which is in accordance

[2] https://spacy.io/models
[3] https://developers.google.com/knowledge-graph
[4] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

Metric	En→ De (w/o Human)			Zh→ En (w/o Human)			En→ Ru (w/o Human)			En→ De (with Human)			Zh→ En (with Human)			En→ Ru (with Human)		
	r	τ	ρ	r	τ	ρ	r	τ	ρ	r	τ	ρ	r	τ	ρ	r	τ	ρ
			WMT 20										**WMT 20**					
BERTScore	0.754	**0.429**	**0.536**	0.742	0.643	0.810	-	-	-	0.281	**0.067**	**-0.018**	0.550	**0.422**	**0.467**	-	-	-
EE-BERTScore-0.3	0.721	**0.429**	**0.536**	**0.896**	**0.714**	**0.833**	-	-	-	**0.297**	-0.067	-0.079	**0.582**	**0.422**	**0.467**	-	-	-
EE-BERTScore-0.5	0.736	**0.429**	**0.536**	0.827	**0.714**	**0.833**	-	-	-	0.292	0.022	-0.030	0.569	**0.422**	**0.467**	-	-	-
EE-BERTScore-0.8	**0.755**	0.333	0.464	0.654	0.571	0.690	-	-	-	0.284	**0.067**	**-0.018**	0.547	0.378	0.406	-	-	-
			WMT 21-news										**WMT 21-news**					
BERTScore	0.911	0.795	**0.945**	0.577	0.308	0.484	0.776	0.538	0.692	0.181	0.441	0.500	0.382	0.295	0.439	0.540	0.417	0.485
EE-BERTScore-0.3	0.874	**0.846**	**0.945**	**0.637**	**0.487**	**0.626**	0.621	0.451	0.622	0.182	**0.485**	0.512	**0.384**	**0.410**	**0.521**	**0.569**	0.317	0.435
EE-BERTScore-0.5	0.898	**0.846**	**0.945**	0.595	0.359	0.511	0.717	0.495	0.701	0.183	0.500	0.517	0.382	0.352	0.457	0.562	0.383	0.491
EE-BERTScore-0.8	**0.919**	0.769	0.923	0.526	0.256	0.462	**0.809**	**0.604**	**0.754**	**0.184**	0.456	**0.532**	0.380	0.276	0.429	0.548	**0.467**	**0.526**
			WMT 21-tedtalks										**WMT 21-tedtalks**					
BERTScore	0.465	0.256	0.319	0.634	0.055	0.134	0.826	0.626	0.793	0.541	0.363	0.455	-0.634	-0.086	-0.079	0.659	0.676	0.832
EE-BERTScore-0.3	**0.560**	0.333	0.473	0.321	0.055	0.125	0.687	0.451	0.626	**0.553**	0.429	0.578	-0.775	-0.086	-0.086	-0.568	0.219	0.289
EE-BERTScore-0.5	0.558	0.333	0.445	0.534	**0.077**	**0.143**	0.750	0.495	0.679	0.549	0.429	0.556	-0.719	**-0.067**	**-0.071**	-0.538	0.276	0.361
EE-BERTScore-0.8	0.495	**0.359**	**0.478**	**0.645**	**0.077**	0.134	**0.829**	**0.692**	**0.829**	0.543	**0.451**	**0.582**	**-0.617**	**-0.067**	-0.079	**0.805**	**0.714**	**0.857**

Table 2: Correlations with system-level human MQM scores on datasets of WMT 20 news, WMT 21 news and WMT 21 tedtalks. EE-BERTScore-$*$ represents EE-BERTScore with different w values. **With Human** indicates evaluation on MT systems and human traslations, and **w/o Human** indicates MT systems only. Best correlations are marked in bold.

with the conclusion from (Freitag et al., 2021b) that most metrics struggle to correctly score translations that are different from MT systems. However, we still see EE-BERTScore-$*$ improves the correlations with human for BERTScore in some cases (En→ De in WMT 21 datasets), while there are cases where EE-BERTScore-$*$ hardly has a difference with BERTScore in terms of the correlations (Zh→ En in WMT 20 news). Overall, when human translations participate as additional outputs, EE metrics bring a less significant improvement to the standard metrics.

3.2 Experiments of Reference-free

This section introduces the experimental results of our four reference-free metrics.

3.2.1 HWTSC-Teacher-Sim

We choose paraphrase-multilingual-mpnet-base-v2[4] as the model for generating sentence embeddings. Triplets were build with source, MT, and the scores of MT - the scores of MT were normalized. The MT with a higher score is closer to the source in the vector space. With TripletEvaluator, we achieve the alignment of embeddings of source and MT in the space vector. In en-de and zh-en, we use MQM data of WMT2020 and WMT2021 as train set and test set respectively. Since en-ru only has MQMdata of WMT2021, the experimental results of en-ru are missing. COMET-QE-DA_2021-src (Rei et al., 2020) is chosen as the state-of-the-art reference-free metric for comparison. And sentBLEU and BLEU (Koehn et al., 2007) are selected as the state-of-the-art reference-based metrics.

The experimental results show that the introduction of multilingual knowledge distillation is more helpful to the system level scoring accuracy of reference-free HWTSC-Teacher-Sim.

Metrics	en-de	zh-en
sentBLEU	0.083	0.176
COMET-QE-DA_2021-src	0.244	0.305
HWTSC-Teacher-Sim	0.205	0.355

Table 3: Segment-level Kendall correlations for language pairs of WMT21 MQM data

Metrics	en-de	zh-en
BLEU	0.937	0.310
COMET-QE-DA_2021-src	0.847	0.453
HWTSC-Teacher-Sim	0.863	0.596

Table 4: System-level Pearson correlations for language pairs of WMT21 MQM data

3.2.2 HWTSC-TLM

XLM-R[5] is selected as the masked language model for our metric HWTSC-TLM. The segment-level and system-level results on the 8 from-English language pairs of WMT19 are reported in Table 5 and Table 6 respectively. YiSi-2 (Lo, 2019) and Prism-src (Thompson and Post, 2020) are chosen as the state-of-the-art unsupervised reference-free metrics for comparison, and reference-based metrics sentBLEU and BLEU (Koehn et al., 2007) are selected for reference. More experimental results of HWTSC-TLM on WMT19 could be found in (Zhang et al., 2022a).

From the results in Table 5 and Table 6, it could be seen that HWTSC-TLM is much better than

[5]https://huggingface.co/xlm-roberta-base

Metrics	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh	Avg
sentBLEU	0.367	0.248	0.396	0.465	0.392	0.334	0.469	0.270	0.368
YiSi-2	0.069	0.212	0.239	0.147	0.187	0.003	-0.155	0.044	0.093
Prism-src	0.470	0.402	0.555	0.215	0.507	0.499	0.486	0.287	0.428
HWTSC-TLM	0.443	0.343	0.492	0.328	0.301	0.471	0.457	0.297	0.392

Table 5: Segment-level metric results for from-English language pairs of WMT19: absolute Kendall's Tau correlation of segment-level metric scores with DA.

Metrics	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh	Avg
BLEU	0.897	0.921	0.969	0.737	0.852	0.989	0.986	0.901	0.907
YiSi-2	0.324	0.924	0.696	0.314	0.339	0.055	0.766	0.097	0.439
Prism-src	0.865	0.976	0.933	0.444	0.959	0.908	0.822	0.793	0.838
HWTSC-TLM	0.896	0.978	0.941	0.683	0.897	0.919	0.819	0.959	0.886

Table 6: System-level metric results for from-English language pairs of WMT19: absolute Pearson correlation of system-level metric scores with DA.

YiSi-2, and is very competitive with Prism-src, which is a very strong baseline in unsupervised reference-free metrics, although only system translations are used in HWTSC-TLM.

3.2.3 CrossQE

Experiments and results of CrossQE could be found in WMT 2022 QE task report (Su et al., 2022).

3.2.4 KG-BERTScore

The ninth layer of XLM-R[5] is selected for word embedding to calculate F_{BERT} scores in our metric KG-BERTScore. The segment-level and system-level results on the 7 into-English language pairs of WMT19 are reported in Table 7 and Table 8 respectively. YiSi-2 (Lo, 2019) and reference-free BERTScore are chosen as unsupervised reference-free metrics for comparison, and reference-based metrics sentBLEU and BLEU (Koehn et al., 2007) are selected for reference. The experimental results show that the introduction of multilingual knowledge graph is more helpful to the system level scoring accuracy of reference-free BERTScore.

Metrics	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en	mean
sentBLEU	0.056	0.233	0.188	0.377	0.262	0.125	0.323	0.223
YiSi-2	**0.068**	0.126	-0.001	0.096	0.075	0.053	**0.253**	0.096
BERTScore	0.036	**0.234**	**0.171**	0.310	**0.211**	0.089	0.196	**0.178**
KG-BERTScore	0.039	0.191	0.165	**0.313**	0.177	**0.095**	0.213	0.170

Table 7: Segment-level metric results for into-English language pairs of WMT19: absolute Kendall's Tau correlation of segment-level metric scores with DA.

4 Conclusions

In this paper, we present one reference-based metric and four reference-free metrics. We apply the methods of entropy-enhance, multilingual knowledge distillation, multilingual knowledge graph, and quality evaluation in MT to WMT 2022 Metrics Shared Task. The experimental results show great effectiveness of our research direction and the superiority of our metrics.

Metrics	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en	mean
BLEU	0.849	0.982	0.834	0.946	0.961	0.879	0.899	0.907
YiSi-2	0.796	0.642	-0.566	-0.324	0.442	-0.339	**0.940**	0.227
BERTScore	0.785	**0.866**	-0.007	0.117	0.657	-0.372	0.728	0.396
KG-BERTScore	**0.862**	0.733	**0.764**	**0.936**	**0.688**	**0.918**	0.908	**0.830**

Table 8: System-level metric results for into-English language pairs of WMT19: absolute Pearson correlation of system-level metric scores with DA.

References

Jeonghun Baek, Yusuke Matsui, and Kiyoharu Aizawa. 2021. What if we only use real datasets for scene text recognition? toward scene text recognition with fewer labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3113–3122.

Yimeng Chen, Chang Su, Yingtao Zhang, Yuxia Wang, Xiang Geng, Hao Yang, Shimin Tao, Guo Jiaxin, Wang Minghan, Min Zhang, Yujia Liu, and Shujian Huang. 2021. HW-TSC's participation at WMT 2021 quality estimation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 890–896, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021a. 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.

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.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jia-Wei Low, Lidong Bing, and Luo Si. 2021. On the effectiveness of adapter-based tuning for pretrained language model adaptation. *arXiv preprint arXiv:2106.03164*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR.

Hyun Kim, Hun-Young Jung, Hongseok Kwon, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator: Neural quality estimation based on target word prediction for machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 17:1–22.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Yilun Liu, Shimin Tao, Chang Su, Min Zhang, Yanqing Zhao, and Hao Yang. 2022. Part represents whole: Improving the evaluation of machine translation system using entropy enhanced metrics. In *Findings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

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.

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.

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.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020a. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020b. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 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.

Chang Su, Miaomiao Ma, Shimin Tao, Hao Yang, Xiang Geng, Shujian Huang, Min Zhang, Jiaxin Guo, Wang Minghan, Min Zhang, et al. 2022. Hw-tsc's participation at wmt 2022 quality estimation shared task. In *Proceedings of the Senventh Conference on Machine Translation*. Submitted.

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

Minghan Wang, Hao Yang, Hengchao Shang, Daimeng Wei, Jiaxin Guo, Lizhi Lei, Ying Qin, Shimin Tao, Shiliang Sun, Yimeng Chen, and Liangyou Li. 2020. HW-TSC's participation at WMT 2020 quality estimation shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1056–1061, Online. Association for Computational Linguistics.

Zhanglin Wu, Min Zhang, Ming Zhu, Yinglu Li, Ting Zhu, Hao Yang, Song Peng, and Ying Qin. 2022. KG-BERTScore: Incorporating Knowledge Graph into BERTScore for Reference-Free Machine Translation Evaluation. In *11th International Joint Conference on Knowledge Graphs, IJCKG2022*. To be publiushed.

Hao Yang, Shimin Tao, Minghan Wang, Min Zhang, Daimeng Wei, Shuai Zhao, Miaomiao Ma, and Ying Qin. 2022a. CCDC: A Chinese-Centric Cross Domain Contrastive Learning Framework. In *Knowledge Science, Engineering and Management*, pages 225–236, Cham. Springer International Publishing.

Hao Yang, Minghan Wang, Daimeng Wei, Hengchao Shang, Jiaxin Guo, Zongyao Li, Lizhi Lei, Ying Qin, Shimin Tao, Shiliang Sun, and Yimeng Chen. 2020. HW-TSC's participation at WMT 2020 automatic post editing shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 797–802, Online. Association for Computational Linguistics.

Hao Yang, Min Zhang, Shimin Tao, Miaomiao Ma, Ying Qin, and Chang Su. 2022b. TeacherSim: Cross-lingual machine translation evaluation with monolingual embedding as teacher. In *The 2nd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*. To be publiushed.

Hui Yu, Xiaofeng Wu, Wenbin Jiang, Qun Liu, and Shouxun Lin. 2015. Improve the evaluation of translation fluency by using entropy of matched sub-segments. *CoRR*, abs/1508.02225.

Min Zhang, Xiaosong Qiao, Hao Yang, Shimin Tao, Yanqing Zhao, Yinlu Li, Chang Su, Minghan Wang, Jiaxin Guo, Yilun Liu, and Ying Qin. 2022a. Target-side language model for reference-free machine translation evaluation. In *The 18th China Conference on Machine Translation, CCMT2022*. To be publiushed.

Min Zhang, Hao Yang, Shimin Tao, Yanqing Zhao, Xiaosong Qiao, Yinlu Li, Chang Su, Minghan Wang, Jiaxin Guo, Yilun Liu, and Ying Qin. 2022b. Incorporating multilingual knowledge distillation into machine translation evaluation. In *The 16th China Conference on Knowledge Graph and Semantic Computing, CCKS2022*. To be publiushed.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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.