Empowering a Metric with LLM-assisted Named Entity Annotation: HW-TSC's Submission to the WMT23 Metrics Shared Task

Zhanglin Wu*, Yilun Liu*, Min Zhang, Xiaofeng Zhao, Junhao Zhu,
Ming Zhu, Xiaosong Qiao, Jingfei Zhang, Miaomiao Ma, Yanqing Zhao
Song Peng, Shimin Tao, Hao Yang, Yanfei Jiang
Huawei Translation Service Center, Beijing, China
{wuzhanglin2,liuyilun3,zhangmin186,zhaoxiaofeng14,zhujunhao,
zhuming47,qiaoxiaosong,zhangjingfei,mamiaomiao,zhaoyanqing,
pengsong2,taoshimin,yanghao30,jiangyanfei}@huawei.com

Abstract

This paper presents the submission of Huawei Translation Service Center (HW-TSC) to the WMT23 metrics shared task, in which we submit two metrics: KG-BERTScore and HWTSC-EE-Metric. Among them, KG-BERTScore is our primary submission for the reference-free metric, which can provide both segment-level and system-level scoring. While HWTSC-EE-Metric is our primary submission for the reference-based metric, which can only provide system-level scoring. Overall, our metrics show relatively high correlations with MQM scores on the metrics tasks of previous years. Especially on system-level scoring tasks, our metrics achieve new state-of-the-art in many language pairs.

1 Introduction

Due to the expensive cost of human evaluation, automatic metrics (Freitag et al., 2022) for machine translation (MT) (Wei et al., 2021, 2022a) is critically important for MT research and development. While human evaluation is still very important, automatic metrics allow the rapid evaluation and comparison of MT systems on large collections of text and facilitate expansion to low resource languages (Li et al., 2022) and domains (Yang et al., 2021; Wu et al., 2022a). Depending on whether the references are required or not, automatic metrics are categorized into two categories: (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 references; (2) reference-free metrics like YiSi-2 (Lo, 2019) and COMET-QE (Rei et al., 2020, 2021), which are also referred to as quality estimation (QE). These metrics estimate the quality of hypothesis based solely on the sources, without relying on the references.

The WMT23 metrics shared task invites submissions of reference-free metrics and reference-based metrics to find automatic metric scores for translations at the segment-level and system-level. This paper presents the contribution of HW-TSC to the WMT23 metrics shared task. Slightly different from our participation last year (Liu et al., 2022a), we only submit two metrics this year. Details of our metrics (KG-BERTScore and HWTSC-EE-Metric) are illustrated in Table 1.

Metric	Category	Segment-level	System-level
KG-BERTScore	reference-free	✓	✓
HWTSC-EE-BERTScore	reference-based	✗	✓

Table 1: Details of our metrics

KG-BERTScore (Wu et al., 2022b) incorporates multilingual knowledge graph (Chen et al., 2017) into BERTScore (Zhang et al., 2019) and generates the final evaluation score by linearly combining the results of KGScore and BERTScore. Our efforts this year build on findings and observations from our participation in the WMT22 metrics shared task (Liu et al., 2022a) to further improve the accuracy of KGScore and BERTScore. The choice of a named entity (NE) annotator (Marrero et al., 2013) is critical to KGScore. With the emergence of large language models (LLMs) (Wei et al., 2022b; Kasneci et al., 2023) such as ChatGPT (Ding et al., 2022), the NE annotator seems to have one more option. Therefore, we try to use ChatGPT[1] for NE annotation and find that LLM-assisted NE annotation can empower the metric. At the same time, the selection of a QE model is crucial for BERTScore. Since COMET-QE (Rei et al., 2022) has proven to be the state-of-the-art QE model, we use it to calculate BERTScore this year.

The HWTSC-EE-Metric (Liu et al., 2022b) is developed using existing metrics with the goal of creating a more balanced scoring system at the sys-

*These authors contributed equally to this work.

[1] https://platform.openai.com

Proceedings of the Eighth Conference on Machine Translation (WMT), pages 822–828
December 6–7, 2023.

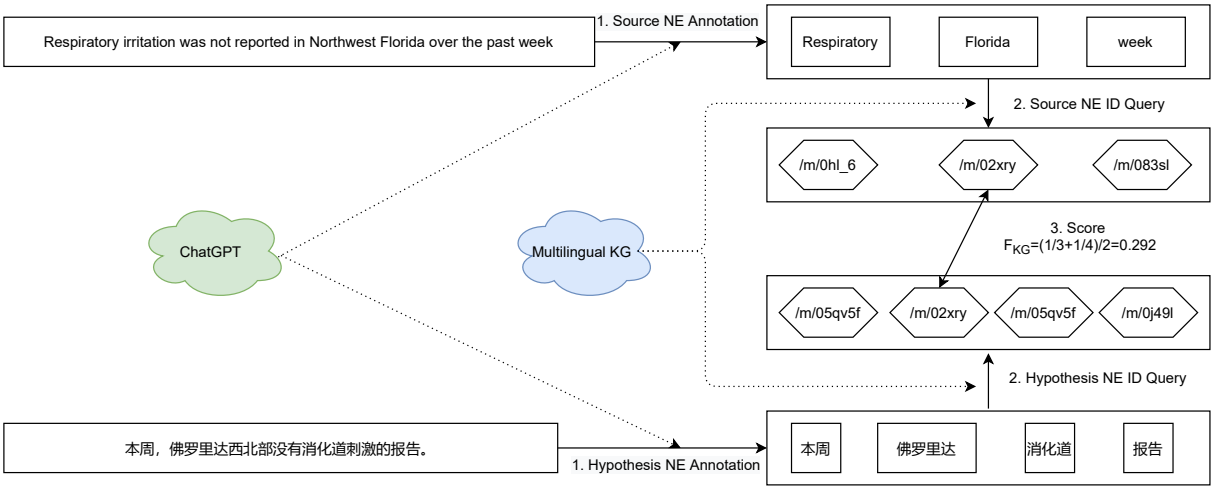

Figure 1: A Calculation Example of KGScore on English-Chinese Language Pair

tem level. This is achieved by assigning weights to segment-level scores obtained from backbone metrics. The weights are determined based on the difficulty of each segment, which is determined by the entropy of a hypothesis-reference pair. Segments with higher entropy values, indicating higher difficulty, receive larger weights in the aggregation of system-level scores by HWTSC-EE-Metric.

2 Metrics

This section introduces our metrics for WMT23 metrics shared task, including KG-BERTScore and HWTSC-EE-Metric.

2.1 KG-BERTScore

KG-BERTScore (Wu et al., 2022b) is a reference-free metric we proposed last year, which generates the final evaluation score by linearly combining the results of KGScore and BERTScore. For a given KGScore F_{KG} and BERTScore F_{BERT}, KG-BERTScore $F_{KG-BERT}$ is defined as:

$$F_{KG-BERT} = \alpha \cdot F_{KG} + (1-\alpha) \cdot F_{BERT}, \quad (1)$$

where α is an adjustable weight parameter.

We have made some improvements to the implementation details of KGScore and BERTScore, which will be described in detail below.

2.1.1 KGScore

KGScore refers to scoring based on the matching rate of NE. Figure 1 is a calculation example of KGScore on English-Chinese language pairs. The calculation process includes three steps:

Firstly, we utilize a NE Annotator to annotate NEs in the source and hypothesis sentences. Last year we used spacy[2] (Algamdi et al., 2022) as the NE annotator, but it didn't work very well. This year we try to use ChatGPT to annotate NE, and find that its effect is better than spacy, which means that LLM-assisted NE annotation is feasible.

Secondly, we match cross-lingual NE pairs by querying multilingual knowledge graphs. Google Knowledge Graph[3] (Google KG) is a general-purpose multilingual knowledge graph that we have chosen to use as always for querying NE IDs. Since same-meaning NEs in different languages share the same NE ID in Google KG, we can match cross-lingual NE pairs by NE ID. One more thing to be noted is that an NE without an ID is considered invalid and will not participate in the subsequent calculation of KGScore.

Finally, we explore using NE's matching rate to score. For a given test set with n sentence pairs, assuming that S_i is the NE numbers in the i-th source sentence, H_i is the NE numbers in the i-th hypothesis sentence, and SH_i is the number of matched cross-lingual NE pairs. The segment-level NE matching rates of the i-th source sentence and hypothesis sentence are respectively defined as:

$$F_{KGS_i} = \frac{SH_i}{S_i} \quad if\ S_i \neq 0\ else\ 1 \quad (2)$$

$$F_{KGH_i} = \frac{SH_i}{H_i} \quad if\ M_i \neq 0\ else\ 1 \quad (3)$$

[2] https://spacy.io/models

[3] https://developers.google.com/knowledge-graph

Then the segment-level calculation formula of KGScore is defined as:

$$F_{KG_i} = \frac{F_{KGS_i} + F_{KGH_i}}{2} \tag{4}$$

For the system-level KGScore, we first calculate the system-level NE matching rates of source sentences and hypothesis sentences are respectively defined as:

$$F_{KGS} = \frac{\sum_{i=1}^{n} SH_i}{\sum_{i=1}^{n} S_i} \tag{5}$$

$$F_{KGH} = \frac{\sum_{i=1}^{n} SH_i}{\sum_{i=1}^{n} H_i} \tag{6}$$

Then the system-level calculation formula of KGScore is defined as:

$$F_{KG} = \frac{F_{KGS} + F_{KGH}}{2} \tag{7}$$

2.1.2 BERTScore

BERTScore (Zhang et al., 2020) refers to scoring based on semantic similarity. We initially use Sentence-BERT (Reimers and Gurevych, 2019) to calculate the semantic similarity score between the source and hypothesis. Last year we used a reference-free HWTSC-teacher-Sim metric (Zhang et al., 2022) as BERTScore to make the score more relevant to MQM score (Lommel et al., 2014). As COMET-QE has been proven to be the state-of-the-art reference-free metric on WMT22 metrics shared task, we use the COMET-QE model[4] to score and serve as BERTScore this year.

2.2 HWTSC-EE-Metric

The HWTSC-EE-Metric, also known as the entropy-enhanced (EE) Metrics (Liu et al., 2022b), was employed in system-level shared tasks this year. Unlike traditional methods of acquiring system-level scores, EE metrics deviate from the normal approach of obtaining system-level scores via arithmetic average. EE metrics assign higher weights to difficult samples present in the evaluation set, as opposed to treating all source-reference pairs equally, as human scorers tend to do in MT evaluation. It is worth noting that simple samples can be easily translated, leading to similar human scores for different hypotheses. Conversely, challenging samples within the evaluation set play a crucial role in differentiating top candidates from inferior systems. Consequently, MT evaluation metrics should encourage systems that excel in translating difficult samples. Contrary to concerns about incorrect scoring, the use of challenging segments to evaluate MT systems has actually shown potential for improving metric performance. EE metrics, in particular, place a strong emphasis on the translation quality of difficult hypotheses and allocate higher weights to them in system-level scores.

2.2.1 Working Process of EE Metrics

EE metrics use the average qualities of hypotheses to determine the difficulty of a segment. One key measure used in this process is chunk entropy (Yu et al., 2015), which quantifies the quality of translation between the reference and the hypothesis. Higher chunk entropy indicates higher uncertainty in translation, while lower entropy suggests good confidence in the hypothesis. By calculating the entropy, easy and difficult samples can be classified accordingly through a threshold value h. In the process of aggregating scores, hypotheses are assigned weights based on their group, whether they belong to the easy or difficult category. Easy samples receive a lower weight denoted as w/N_e, while difficult samples receive a higher weight $(1-w)/N_d$. The reason for such a weight discrepancy lies in the larger number of easy hypotheses compared to difficult ones. The balance coefficient w may vary depending on the language pairs and evaluation datasets utilized. This weight assignment strategy ensures that the weights of easy samples remain significantly lower than those of difficult samples, considering the different samples in each category.

2.2.2 Enhancements to HWTSC-EE-Metric

The earlier version of EE metrics incorporates two adjustable hyperparameters, h and w, which are responsible for selecting difficult samples and assigning weights to each group, respectively. However, the presence of these hyperparameters hampers the practical application of EE metrics. Furthermore, these hyperparameters often vary across different language pairs and evaluation datasets, as evidenced by our preliminary experiment that involved up to 10 different parameters using the WMT19 evaluation set. Consequently, it becomes challenging to identify a suitable combination of hyperparameters for real-world scenarios. To address this issue, in last year's WMT metrics shared tasks (Liu et al., 2022a), we simplified the com-

[4] https://huggingface.co/Unbabel/wmt22-cometkiwi-da

putation of the system-level score by employing a normal distribution fitting approach to determine the threshold h for each translation direction. This year, we further simplified the estimation of w by using a fixed value of 0.8, as opposed to the three different configurations of w used last year. Based on the results of WMT22, we observed that the value 0.8 corresponds to an appropriate balance of weights between difficult and easy groups, as it exhibits a high correlation with human MQM scores on recent WMT test sets. Another modification in this year's HWTSC-EE-Metric is the replacement of our backbone metric from BERTScore (Zhang et al., 2019) to COMET score (Rei et al., 2022). Specifically, we adopted the model *wmt22-comet-da*[5], which is known for its robust segment-level MT evaluation capabilities, as the segment-level backbone metric for HWTSC-EE-Metric this year.

3 Experiments

This section introduces the experimental results of KG-BERTScore and HWTSC-EE-Metric on previous metrics shared tasks.

3.1 Experiment of KG-BERTScore

In order to verify the feasibility of the improved KG-BERTScore, we conduct experiments on the WMT22 metrics shared task data. Since it is time-consuming and expensive to query ChatGPT and Google Knowledge Graph API, we only verify the effect of KG-BERTScore on Chinese-English language pair. In the experiment, we first calculate F_{KGS} and F_{KGH} through NE annotation and NE pair matching, and then calculate KGScore. Next, we use COMETKiwi-22 as BERTScore to calculate the final KG-BERTScore.

We calculate the correlation of the scores of each stage (including F_{KGS}, F_{KGH}, KGscore and KG-BERTScore) with the MQM scores without considering human translation. To facilitate comparison with the official results of the WMT22 metrics shared task, the segment-level correlation adopts Kendall correlation, and the system-level correlation adopts Pearson correlation.

3.1.1 Segment-level Correlation

Table 2 shows Kendall Tau correlation of reference-free metrics with segment-level MQM scores for the WMT22 Chinese-English language pair, which is calculated without human translation. We find that KGScore has a relatively low segment-level correlation with MQM scores, while COMETKiwi-22 has a relatively high segment-level correlation with MQM scores. Therefore, when calculating KG-BERTScore, we set α to a smaller value (i.e., 0.1). Overall, the segment-level correlation between KG-BERTScore and MQM scores is only slightly higher than that of COMETKiwi-22.

Metric	Correlation
KG-BERTScore-22	0.219
COMETKiwi-22	0.364
F_{KGS}	0.017
F_{KGH}	0.055
KGScore	0.061
KG-BERTScore (α=0.1)	**0.365**

Table 2: Kendall Tau correlation of reference-free metrics with segment-level MQM scores for the WMT22 Chinese-English language pair, which is calculated without human translation.

3.1.2 System-level Correlation

Table 3 shows Pearson correlation of reference-free metrics with system-level MQM scores for the WMT22 Chinese-English language pair, which is calculated without human translation. The system-level correlation between KGScore and MQM scores is relatively close to that of COMETKiwi-22, so we set α to a larger value (i.e., 0.9). Surprisingly, the system-level correlation between KG-BERTScore and MQM scores is significantly higher than that of COMTKiwi-22.

In addition, the segment-level and system-level correlations of KGScore with MQM scores are higher than those of F_{KGS} and F_{KGH}, which indicates that both source and hypothesis NE pair matching rates should be considered when calculating KGScore.

Metric	Correlation
KG-BERTScore-22	0.743
COMETKiwi-22	0.866
F_{KGS}	0.660
F_{KGH}	0.376
KGScore	0.697
KG-BERTScore (α=0.9)	**0.947**

Table 3: Pearson correlation of reference-free metrics with system-level MQM scores for the WMT22 Chinese-English language pair, which is calculated without human translation.

[5]https://huggingface.co/Unbabel/wmt22-comet-da

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
WMT21-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
COMET	0.812	0.590	0.819	0.545	0.359	0.401	0.774	0.538	0.688	0.349	0.559	0.804	0.425	0.333	0.386	0.751	0.617	0.782
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
HWTSC-EE-Metric	0.816	0.615	0.819	0.474	0.359	0.462	**0.814**	0.582	0.727	**0.380**	**0.574**	**0.806**	**0.427**	0.333	0.454	**0.761**	**0.683**	**0.821**
WMT21-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
COMET	0.764	0.436	0.604	0.620	**0.143**	0.196	**0.878**	0.692	**0.868**	0.626	0.516	0.684	-0.638	**-0.010**	-0.029	0.784	**0.733**	**0.893**
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
HWTSC-EE-Metric	**0.799**	**0.538**	**0.742**	0.633	**0.143**	**0.213**	0.869	**0.851**	0.692	**0.653**	**0.604**	**0.793**	**-0.593**	**-0.010**	**-0.014**	-0.005	0.467	0.504

Table 4: Correlations with system-level human MQM scores on datasets of WMT21 news and WMT21 tedtalks. EE-BERTScore-$*$ represents our last year's submission in WMT22. HWTSC-EE-Metric represents our submission in WMT23. **With Human** indicates evaluation on MT systems and human translations, and **w/o Human** indicates MT systems only. Best correlations are marked in bold.

3.1.3 Effect of Different Weights

KG-BERTScore generates the final evaluation score by linearly combining the results of KGScore and BERTScore. α is an adjustable weight parameter in the linear combination formula, which affects the correlation between KG-BERTScore and MQM scores. To analyze the effect of α value, we calculate the segment-level and system-level correlations of KG-BERTScore and MQM scores under different α values for the WMT22 Chinese-English language pair. The result is shown in Figure 2. The segment-level correlation between KG-BERTScore and MQM scores is highest when the α value is 0.1, and the system-level correlation between KG-BERTScore and MQM scores is the highest when the α value is 0.9. That is to say, when the correlation between KGScore and MQM scores is relatively low, α should take a smaller value, otherwise, α should set a larger value.

On the WMT23 metrics shared task, we cannot know the MQM score in advance. Therefore, we refer to the above experimental settings, and set α to 0.1 and 0.9 on the segment-level and system-level metrics shared tasks, respectively. In addition, we do not calculate KG-Score and set α to 0 on non-MQM language pairs due to the slow speed of accessing ChatGPT.

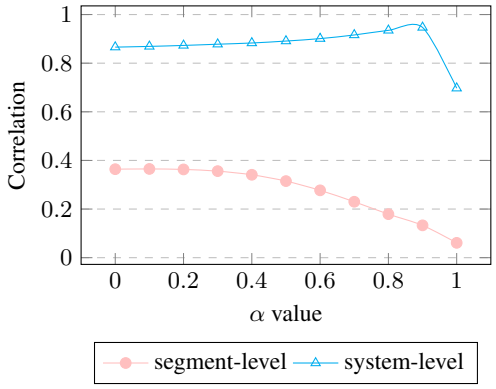

Figure 2: The segment-level and system-level correlations between KG-BERTScore and MQM scores under different α values for the WMT22 Chinese-English.

3.2 Experiment of HWTSC-EE-Metric

To evaluate the performance of the HWTSC-EE-Metric, a series of experiments were conducted primarily on the WMT21 test sets using the MQM scores as the human scoring standard. To investigate the impact of using human translations as part of the system, the results obtained from two sets of systems for each language pair are compared. The evaluation was based on three coefficients: Pearson's correlation coefficient (r), Kendall's τ, and Spearman's ρ, which are used to assess the system-level correlations with human evaluations.

Table 4 presents a performance comparison between the HWTSC-EE-Metric (our submission in WMT23), EE-BERTScore (our submission in WMT22), and two standard metrics (BERTScore and COMET). The HWTSC-EE-Metric demonstrates higher overall correlations with human MQM evaluations compared to its backbone, the standard COMET score. Furthermore, out of the 36 comparison terms, the HWTSC-EE-Metric achieves the best performance in 20 cases. This strong performance indicates the effectiveness of our entropy-based enhancing strategy and parameter estimation approach.

As EE metrics evaluate a system based not only on the individual system itself but also on other participating systems, the inclusion of human trans-

lations may influence the performance of EE metrics. As shown in Table 4, most metrics exhibit a decline in performance when human translations are included. The improvements of the HWTSC-EE-Metric in correlations with MQM are not consistently steady, which aligns with the findings of (Freitag et al., 2021) that most metrics struggle to accurately score translations that differ from MT systems. However, we observed that the HWTSC-EE-Metric mitigates the performance reduction of COMET in some cases (*e.g.*, En→ De in WMT21 datasets), but there are also instances where the HWTSC-EE-Metric does not improve COMET in terms of correlations (*e.g.*, En→ Ru in WMT21 TED talks). Overall, when human translations are included as additional outputs, EE metrics tend to be less robust and provide a less significant improvement over standard metrics.

4 Conclusion

This paper presents HW-TSC's submission to the WMT23 metrics shared task, in which we summit a reference-free metric (KG-BERTScore) and a reference-based metric (HWTSC-EE-Metric). We have made some improvements to these two metrics compared to last year's submission. One of the most critical improvements is on KG-BERTScore, we empower the metric with LLM-assisted NE annotations, significantly improving its correlation with MQM scores. The experimental results on previous WMT metrics tasks show great effectiveness of our research direction and the superiority of our metrics.

References

Shabbab Algamdi, Abdullah Albanyan, Sayed Khushal Shah, and Zeenat Tariq. 2022. Twitter accounts suggestion: Pipeline technique spacy entity recognition. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5121–5125. IEEE.

Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1511–1517.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq R. Joty, and Boyang Li. 2022. Is gpt-3 a good data annotator? In *Annual Meeting of the 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 FT 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.

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

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

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.

Shaojun Li, Yuanchang Luo, Daimeng Wei, Zongyao Li, Hengchao Shang, Xiaoyu Chen, Zhanglin Wu, Jinlong Yang, Zhiqiang Rao, Zhengzhe Yu, et al. 2022. Hw-tsc systems for wmt22 very low resource supervised mt task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1098–1103.

Yilun Liu, Xiaosong Qiao, Zhanglin Wu, Chang Su, Min Zhang, Yanqing Zhao, Song Peng, Shimin Tao, Hao Yang, Ying Qin, et al. 2022a. Partial could be better than whole. hw-tsc 2022 submission for the metrics shared task. *WMT 2022*, page 549.

Yilun Liu, Shimin Tao, Chang Su, Min Zhang, Yanqing Zhao, and Hao Yang. 2022b. Part represents whole: Improving the evaluation of machine translation system using entropy enhanced metrics. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 296–307.

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. *Tradumàtica*, pages 0455–463.

Mónica Marrero, Julián Urbano, Sonia Sánchez-Cuadrado, Jorge Morato, and Juan Miguel Gómez-Berbís. 2013. Named entity recognition: fallacies,

challenges and opportunities. *Computer Standards & Interfaces*, 35(5):482–489.

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, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

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)*. 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.

Daimeng Wei, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, et al. 2021. Hw-tsc's participation in the wmt 2021 news translation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 225–231.

Daimeng Wei, Zhiqiang Rao, Zhanglin Wu, Shaojun Li, Yuanchang Luo, Yuhao Xie, Xiaoyu Chen, Hengchao Shang, Zongyao Li, Zhengzhe Yu, et al. 2022a. Hw-tsc's submissions to the wmt 2022 general machine translation shared task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 403–410.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022b. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Zhanglin Wu, Jinlong Yang, Zhiqiang Rao, Zhengzhe Yu, Daimeng Wei, Xiaoyu Chen, Zongyao Li, Hengchao Shang, Shaojun Li, Ming Zhu, et al. 2022a. Hw-tsc translation systems for the wmt22 biomedical translation task. In *Proceedings of the Seventh Conference on Machine Translation*, pages 936–942.

Zhanglin Wu, Min Zhang, Ming Zhu, Yinglu Li, Ting Zhu, Hao Yang, Song Peng, and Ying Qin. 2022b. 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, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Daimeng Wei, Zongyao Li, Hengchao Shang, Minghan Wang, Jiaxin Guo, Lizhi Lei, et al. 2021. Hw-tsc's submissions to the wmt21 biomedical translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 879–884.

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

Min Zhang, Hao Yang, Shimin Tao, Yanqing Zhao, Xiaosong Qiao, Yinlu Li, Chang Su, Minghan Wang, Jiaxin Guo, Yilun Liu, and Ying Qin. 2022. 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.