Part Represents Whole: Improving the Evaluation of Machine Translation System Using Entropy Enhanced Metrics

Yilun Liu, Shimin Tao*, Chang Su, Min Zhang, Yanqing Zhao, Hao Yang
Huawei Translation Services Center, Beijing, China
{liuyilun3,taoshimin,suchang8,zhangmin186,
zhaoyanqing,yanghao30}@huawei.com

Abstract

Machine translation (MT) metrics often fail to achieve very high correlations with human assessments. In terms of MT system evaluation, most metrics pay equal attentions to every sample in an evaluation set, while in human evaluation, difficult sentences often make candidate systems distinguishable via notable fluctuations in human scores, especially when systems are competitive. We find that samples with high entropy values, which though usually count for less than 5%, tend to play a key role in MT evaluation: when the evaluation set is shrunk to only the high-entropy portion, correlations with human assessments are actually improved. Thus, in this paper, we propose a fast and unsupervised approach to enhance MT metrics using entropy, expanding the dimension of evaluation by introducing sentence-level difficulty. A translation hypothesis with a significantly high entropy value is considered difficult and receives a large weight in aggregation of system-level scores. Experimental results on five sub-tracks in the WMT19 Metrics shared tasks show that our proposed method significantly enhanced the performance of commonly-used MT metrics in terms of system-level correlations with human assessments, even outperforming existing SOTA metrics. In particular, all enhanced metrics exhibit overall stability in correlations with human assessments in circumstances where only competitive MT systems are included, while the corresponding standard metrics fail to correlate with human assessments[1].

1 Introduction

Automatic evaluation plays an indispensable role in the evaluation of machine translation (MT) systems, working as a proxy of human assessment as well as a promising approach to give instant feedback during the development of MT systems. However, it has been a challenge for automatic evaluations to correlate with human judgement. For instance, major discrepancy is detected between human assessments and automatic evaluations in terms of system ranking in WMT19 English-German evaluation tasks (Barrault et al., 2019). Experiments conducted by Mathur et al. (2020) and Thompson and Post (2020) further indicate that when inferior systems are excluded, current automatic metrics expect major falling on correlations with human referees, sometimes even down to the degree of negative correlations.

In order to improve the evaluation of MT systems, many meticulously designed metrics are proposed. However, popular MT metrics focus on a segment-level comparison between references and hypotheses, and output system-level scores by a simple arithmetic average over segment scores, ignoring the differences among samples in an evaluation set (Zhang et al., 2019; Sellam et al., 2020; Rei et al., 2020; Lo, 2020). In contrast, the core idea of assigning different weights to samples in a dataset is proven effective in the field of curriculum learning (Liu et al., 2020; Zhan et al., 2021b). 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. Inspired by recent work of Zhan et al. (2021a), who determine the difficulty of sub-units in translation hypotheses by reviewing performances of corresponding sub-units among K candidate systems, we further introduce sentence-level difficulty into MT evaluation, which functions as a weight in the aggregation of final system scores. In determination of proposed sentence-level difficulty, instead of using an embedding-based approach similar to Zhan et al.'s, we adopt a fast and unsupervised entropy-based measurement.

* Corresponding author
[1]Code at https://github.com/lunyiliu/EE-Metrics

Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, pages 296–307
November 20–23, 2022.

In information theory, entropy is a measure of the uncertainty in a random variable. The entropy H of a discrete random variable X with possible values $x_1, x_2, ..., x_n$ is defined by Shannon (1948) as

$$H(X) = -\sum_{i=1}^{n} P(x_i) log_2 P(x_i), \quad (1)$$

where $P(x_i)$ is the probability for x_i to appear in the stream of characters. The entropy $H(X)$ will be higher if the values $x_1, x_2, ..., x_n$ are more decentralized. So the entropy can reflect the degree of disorder of variable X's distribution. Shannon's standard entropy is interpreted differently when being applied to MT evaluation (Zhao et al., 2019; Yu et al., 2015). Zhao et al. (2019) define x_i in Eq.(1) as the ith candidate among all possible translations of a source token X, while Yu et al. (2015) directly model one hypothesis produced by a system as random variable X and consider x_i as the ith sub-segment in the hypothesis matched with corresponding reference sentence. We follow the idea of chunk entropy in Yu et al. (2015). Compared with token difficulty in Zhan et al. (2021a), which requires a loop of K systems' hypotheses for each token, chunk entropy can determine the difficulty of hypotheses in constant time, reflecting both *adequacy* and *fluency* of a hypothesis. This will be further discussed in section 3.

In this paper, we propose entropy enhanced (EE) metric, a criterion that can enhance the performances of automatic MT metrics via a sentence-level translation difficulty weight determined by entropy. The difficulty score of each hypothesis-reference pair is acquired based on its chunk entropy and then serves as a weight in aggregation of the system-level score. Experiments carried on WMT19 evaluation tasks show that the EE version of BERTScore (Zhang et al., 2019) correlates better with system-level human ratings than DA-BERTScore (Zhan et al., 2021a) and outperforms SOTA metrics involved in WMT metrics shared tasks. Also, owing to the sentence-level difficulty dimension and the underlying essence of entropy, the proposed method should be compatible with a wide range of MT evaluation metrics. We test the effectiveness on several representative metrics in addition to BERTScore: BLEU (Papineni et al., 2002), CHRF (Popović, 2015) and METEOR (Denkowski and Lavie, 2014). Extensive experiments on five sub-tracks in WMT19 indicate an overall improvement on correlations with human evaluations when standard metrics are replaced by corresponding EE metrics. Moreover, in circumstances where only competitive systems are included, EE metrics alleviate the significant crash of standard metrics on correlations, and sometimes even achieve perfect agreements with human rankings.

It is surprising to see a straightforward implementation under the idea of sentence-level difficulty weights based on entropy, involving no deep-learning techniques, yet enhanced the performance of a BERT-based MT metric. The aim of this paper is to introduce the concepts and show the effective roles entropy and sentence-level difficulty play in enhancing MT evaluation quality, but not to explore optimal techniques integrating them into MT evaluation.

2 Related Work

Existing reference-based MT metrics can be roughly categorized into three types: matching-based metrics (Doddington, 2002; Papineni et al., 2002; Popović, 2015; Snover et al., 2006; Leusch et al., 2006; Denkowski and Lavie, 2014), embedding-based metrics (Zhang et al., 2019; Chow et al., 2019; Lo, 2019) and end-to-end metrics (Sellam et al., 2020; Rei et al., 2020). Matching-based metrics estimate quality of translation by hand-crafted features, such as n-grams, edit distance and alignments. BLEU (Papineni et al., 2002) is a classical criterion based on word-level n-gram matching between references and hypothesis and is widely employed as baselines in MT system evaluation, while CHRF (Popović, 2015) computes an F-score based on character-level n-grams. METEOR (Denkowski and Lavie, 2014) focuses on semantic matched chunks acquired by alignment, where lengths of chunks are dynamically determined and the limitation of maximum matching length of n-gram based metrics is partially relieved. In contrast, BERTScore and its variants (Zhang et al., 2019; Zhan et al., 2021a), owing to powerful contextual embedding acquired from modern language models, catch deep-level semantic information inside the translation pairs and achieve high rankings across MT evaluation benchmarks in terms of correlations with human assessments.

3 Our Proposed Method

3.1 Motivation

In the evaluation of MT systems, most automatic metrics rate a system by the average scores on sentences in the evaluation set, treating each segment equally, while assigning weights to samples has been successful in the practice of curriculum learning (Liu et al., 2020). Like examinations in real world, where questions are assigned different weights in the final score based on variant difficulties, evaluation metric of MT should also encourage systems that perform better on relatively difficult samples. Also, in competitive circumstance where candidates can handle most of the easy translations, difficult samples can better represent the abilities of candidates. In contrast to (Zhan et al., 2021a), where they compute the difficulty of each sub-unit inside a hypothesis, we directly assign different weights to high-entropy and low-entropy hypotheses so that the more difficult translations weight higher in the final system score.

When entropy is higher, the translation is faced with more uncertainty, leading to potential blemish in *adequacy* and *fluency*. Motivated by this mechanism, we use entropy as a measurement of sentence-level difficulty. Empirically, we found that there is a high negative correlation between entropy and BLEU score of a translation, as shown in Fig. 1. The linear fit shows that BLEU score exhibits a linear decline when entropy increases, with $|r| = 0.986$. When a certain source sentence is difficult to translate, the quality of generated hypothesis may be affected, causing a relatively low average BLEU score. So the difficult samples in an MT evaluation set tend to appear in the high-entropy area, and should be assigned a higher weight in the assessment.

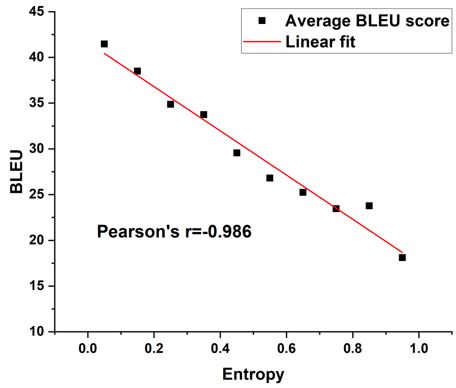

Figure 1: Average sentence-level BLEU score as a function of entropy. Each data point (e, b) represents mean BLEU across sentences with entropy in a range of $[e - 0.05, e + 0.05)$ among outputs of all 22 systems in WMT19 English→German evaluation set.

3.2 Entropy Enhanced MT Metric

In this section, we illustrate the working process of the proposed EE method. As shown in Fig. 2, first, entropy of each hypothesis (H) is calculated and guides the computation of the difficulty weight (W). Then, in aggregation of the final score, W is assigned to the corresponding hypothesis, weighting its sentence-level score.

Chunk Entropy Entropy measures uncertainty or disorderness of the distribution of a variable. In machine translation, a hypothesis generated from a source can be modeled as a random variable $X_h = \{w_1, w_2, ..., w_N\}$ with w_i ($i \in [1...N]$) denoting each token in the hypothesis. Given a reference $R = \{r_1, r_2, ..., r_M\}$, X_h can be rewritten as $X_h = x_1 \cdot u_1 \cdot x_2 \cdot u_2 \cdot ... \cdot u_m \cdot x_n$, where $x_i \in X = \{w_{s_i}, w_{s_i+1}, ..., w_{e_i} \mid i \in [1...n], 1 \le s_i \le e_i \le N, \forall\ l \in [s_i, e_i], w_l \in R\}$, and $u_i \in U = \{w_{b_i}, w_{b_i+1}, ..., w_{o_i} \mid i \in [1...m], 1 \le b_i \le o_i \le N, \forall\ l \in [b_i, o_i], w_l \notin R\}$. In other words, x_i denotes the ith continuously matched chunk with reference, while U denotes unmatched parts between aligned chunks. Since X and U are complementary, the distribution of X_h can be fully described by

$$P(x_i) = \frac{e_i - s_i + 1}{\sum_{j=1}^{n}(e_j - s_j + 1)}\ , \qquad (2)$$

where $x_i \in X$ and s_i, e_i represent the start index and end index of the ith matched chunk, respectively. By substituting Eq. (2) into Eq. (1), we obtain the formula of chunk entropy (Yu et al., 2015)

$$H(X_h) = -\sum_{i=1}^{n} \frac{e_i - s_i + 1}{\sum_{j=1}^{n}(e_j - s_j + 1)} log(\frac{e_i - s_i + 1}{\sum_{j=1}^{n}(e_j - s_j + 1)}) \quad (3)$$

From Eq. (3), when a hypothesis is perfectly matched with corresponding reference, $P(x_i)$ from Eq. (2) is always 1 since there is only one chunk x_1, leading to a zero chunk entropy. Another corner case is that, when there is no token in common between the hypothesis and the reference, there is no matched chunk. In this case, we define $P(x_i)$ as 0 and the entropy approaches positive infinity, suggesting no certainty at all. In practice, a machine generated hypothesis often fails to preserve

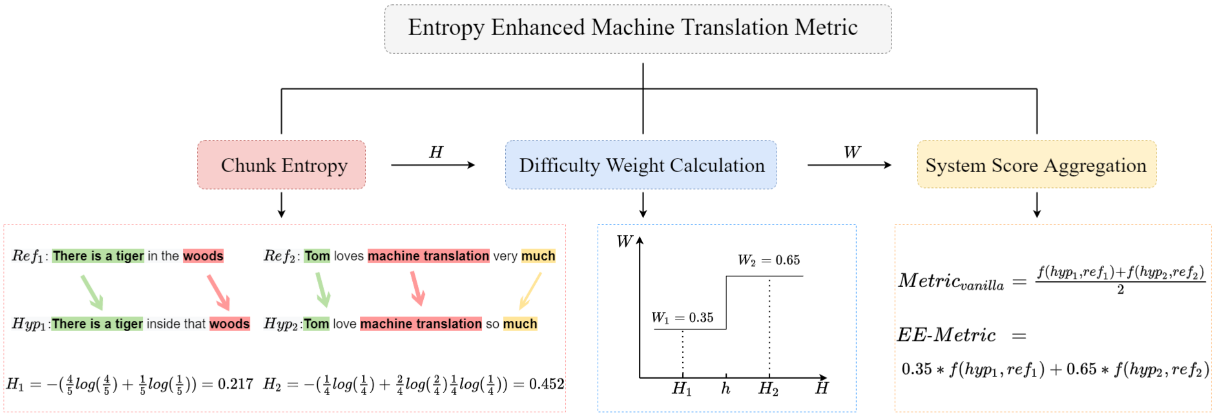

Figure 2: Workflow of proposed entropy enhancement method. $Metric_{standard}$ denotes the system-level score given by a standard MT metric with $f(\cdot)$ as the corresponding sentence-level score function, while $EE\text{-}Metric$ denotes system score aggregated by the corresponding EE metric.

full meaning of the source sentence, or suffers disfluency in the target language (Banchs et al., 2015). Table 1 shows two cases of ascended entropy caused by deficiencies in *adequacy* or *fluency*. The mistranslated word *sheep* in hypothesis 1 sharply increases entropy, while the incorrect word order in hypothesis 2 further deviates the entropy.

	Sentence	Deficiency	Entropy
Reference	A tiger stays in the woods	-	0
Hypothesis 1	**A** sheep **stays in the woods**	*adequacy*	0.217
Hypothesis 2	**A stays** sheep **in the woods**	*adequacy+fluency*	0.292

Table 1: Toy examples of how defect in *adequacy* and *fluency* may lead to increment in entropy of a translation. The matched words in hypotheses are in bold.

Difficulty Weight Calculation With the increasing of entropy, a segment might be faced with more fluctuations in human scores and tends to be representative of quality of systems. Thus, for a certain system, all its generated hypotheses can be divided into the difficult part and easy part by a threshold value of entropy. Those difficult hypotheses are most likely to reflect the ability of a system and distinguish performances among systems, and thus should be weighted higher than those in the easy part. Based on this idea, given $\chi_S = \{X^S_{h_1}, X^S_{h_2}, ..., X^S_{h_L}\}$ as the collection of hypotheses produced by system S in an evaluation set containing L segments, the difficulty weight function can be defined as a two-piece step function:

$$W(H) = \begin{cases} \dfrac{w}{N_e}, & H < h \\ \dfrac{1-w}{N_d}, & H \geq h, \end{cases} \tag{4}$$

where $N_e = |\chi_e|$ and $N_d = |\chi_d|$ are two normalization factors representing the number of easy and difficult hypotheses, respectively, with $\chi_e = \{X_{h_k} \mid H(X_{h_k}) < h, \forall X_{h_k} \in \chi_S\}$ and $\chi_d = \{X_{h_k} \mid H(X_{h_k}) \geq h, \forall X_{h_k} \in \chi_S\}$. And w is a balance coefficient ranging from 0 to 1, and h is the difficulty threshold.

In Eq. (4), h can be defined as the minimal entropy of a generally difficult translation among P systems $S_1, S_2, ..., S_P$. Let X_{s_k} be the source sentence of the kth sample in the evaluation set and $\hat{X}_{s_k} = \{X^{S_1}_{h_k}, X^{S_2}_{h_k}, ..., X^{S_P}_{h_k}\}$ be the collection of translation hypotheses all P systems produced. For system S_p, if $X^{S_p}_{h_k} \in \hat{X}_{s_k}$ has significantly high entropy among other hypotheses in $\hat{X}_{s_k}$, it is reasonable to doubt the quality of hypothesis $X^{S_p}_{h_k}$ and conclude that the source sentence X_{s_k} might be a difficult sample for system S_p. In contrast, when $\overline{H}_{\hat{X}_{s_k}}$ (the average entropy of hypotheses in $\hat{X}_{s_k}$) is significantly higher than that of hypotheses from other source sentences, source X_{s_k} becomes a generally difficult sample. For such a group of source sentences, the minimum value of average entropy among them is actually a threshold to classify easy hypotheses and difficult hypotheses, namely,

$$h = \min\{\overline{H}_{\hat{X}_{s_i}} \mid P(\overline{H}_{\hat{X}_{s_i}} < \overline{H}_{\hat{X}_{s_j}}) < \alpha, \forall i, j \in [1, L], j \neq i\}, \tag{5}$$

where α is a small constant, i.e., 0.05 or 0.01. So the collection of general difficult source sentences can be defined as $D_s = \{X_{s_k} \mid \forall k \in [1, L], \overline{H}_{\hat{X}_{s_k}} \geq h\}$.

From Eq. (5), we can see that the number of easy samples, i.e., when $H < h$, should be larger than the number of difficult ones. So in Eq. (4), we have

Metric	En→ De			De→ En			En→ Zh			Zh→ En			En→ Gu		
	r	τ	ρ	r	τ	ρ	r	τ	ρ	r	τ	ρ	r	τ	ρ
BLEU	0.959	0.755	0.904	0.890	0.655	0.825	0.713	0.606	0.755	0.888	0.695	0.857	0.736	0.709	0.864
CHRF	0.983	0.772	0.919	0.917	0.639	0.822	0.822	0.545	0.650	0.952	0.714	0.868	0.851	0.709	0.891
METEOR	0.986	0.764	0.917	0.837	0.571	0.763	0.513	0.455	0.594	0.946	0.752	0.882	0.820	0.673	0.836
BERTScore	0.990	0.807	0.931	0.954	0.756	0.890	0.909	0.667	0.776	0.986	0.829	0.932	0.902	**0.818**	**0.945**
ESIM	0.991	-	-	0.941	-	-	0.931	-	-	0.988	-	-	-	-	-
YiSi-1	0.991	-	-	0.949	-	-	0.951	-	-	0.979	-	-	0.909	-	-
DA-BERTScore	0.991	0.798	0.930	0.951	0.807	0.932	-	-	-	-	-	-	-	-	-
EE-BLEU	0.965	0.772	0.913	0.882	0.740	0.872	0.727	0.697	0.797	0.907	0.733	0.875	0.787	0.709	0.873
EE-CHRF	0.983	0.798	0.933	0.894	0.639	0.770	0.831	0.545	0.706	0.965	0.752	0.900	0.886	0.745	0.909
EE-METEOR	0.987	0.816	0.940	0.792	0.706	0.854	0.611	0.545	0.636	0.951	0.810	0.936	0.884	0.636	0.836
EE-BERTScore	**0.994**	**0.859**	**0.952**	**0.956**	**0.840**	**0.947**	**0.952**	**0.818**	**0.888**	**0.989**	**0.905**	**0.975**	**0.939**	**0.818**	**0.945**

Table 2: Correlations with system-level human assessments on WMT19 metrics shared task. Best correlations in each column are highlighted in bold. The dashed line separates proposed EE metrics from others. Correlations of DA-BERTScore are directly from Zhan et al. (2021a), and ESIM, YiSi-1 from Ma et al. (2019). Numbers of participated systems for each language pairs are 22, 16, 12, 15 and 11, respectively.

$N_e \gg N_d$, which means simpler samples receive an extremely lower weight than difficult samples. Ideally, the value of $W(H)$ should only be determined by the average entropy of the difficult or simple sample group. To alleviate the distortion caused by unbalanced size between the difficult group and easy group, w, as shown in Eq. (4), is introduced as a balancing coefficient, and can be estimated by the distribution of average entropy within a given dataset. See more analysis on w in appendix B.

System Score Aggregation The designations of most automatic MT metrics focus on the segment level. When outputting system-level ratings, a conventional approach is to aggregate segment-level scores via simple arithmetic averaging. In contrast, the proposed EE metric, when computing system-level scores, assigns a normalized weight, computed by Eq. (4), to the score of each segment. Let $f(\cdot)$ be the unit score function, and the final score is given by

$$EE\text{-}Metric = \textstyle\sum_{i=1}^{L}(W(H(X_{h_i})) \cdot f(X_{h_i}, R_i)), \quad (6)$$

where $H(X_{h_i})$, the chunk entropy of the ith translation, is determined by Eq. (3). For standard metrics, the weight $W(H(X_{h_i}))$ is constantly $1/L$.

In cases where a metric outputs a system-level score based on a whole set of sentences with no segment-level scores involved, i.e., system-level score is directly given by $f(\chi_S)$, an alternative form of EE metric can be obtained via an equivalent transform of Eq. (6):

$$EE\text{-}Metric = wf(\chi_e) + (1-w)f(\chi_d) \quad (7)$$

4 Experiments

Data We follow the experiment settings in Zhan et al. (2021a) for the convenience of comparison and evaluate the performance of EE metrics on WMT19 English↔German (En↔De) evaluation tasks, which is reported to be challenging due to major discrepancy between human assessments and automatic metrics in MT system ranking (Freitag et al., 2020; Barrault et al., 2019). Extended experiments on WMT19 English↔Chinese and English→Gujarati are also conducted to further validate the effectiveness of the proposed approach on both high-resource (En↔Zh) and low-resource (En→Gu) languages, without loss of generality. For every translation task, human ratings of participated systems, in the form of Direct Assessment (DA), are given and the goal of the experiment is to correlate with system-level human DA. Human assessors are asked to rate a given translation by how adequately it expresses the meaning of the corresponding reference translation or source language input on a rating scale of 0-100 (Barrault et al., 2019). For each translation task, there are 21523 assessments and 1592 assessments per participated system in average, given by a total of 1706 crowd-sourced workers. For the sake of quality control, about 20% of the efforts are wasted. Overall, the reliability of human annotators is still relatively high, with the lowest language pair still reaching 88% of workers showing no significant difference in scores for repeat assessment of the same translation.

Comparing Metrics To examine the universal feasibility of the proposed method, we employ four most commonly used MT evaluation metrics as backbones to implement corresponding EE met-

Metric / EE-Metric	En→ De (Top 4)			De→ En (Top 4)			Zh→ En (Top 4)			Average (× 100%)		
	r	τ	ρ	r	τ	ρ	r	τ	ρ	Δr	$\Delta \tau$	$\Delta \rho$
BLEU / EE-BLEU	-0.946 / -0.980	-0.667 / -0.667	-0.800 / -0.800	-0.787 / -0.341	-0.548 / -0.183	-0.632 / -0.316	-0.675 / 0.416	-0.333 / 0.333	-0.600 / 0.400	**+50.10%**	**+34.37%**	**+43.87%**
CHRF / EE-CHRF	-0.677 / 0.013	-0.667 / -0.333	-0.800 / -0.400	-0.659 / -0.240	-0.548 / -0.183	-0.632 / -0.316	-0.353 / 0.657	0.000 / 0.667	0.000 / 0.800	**+70.63%**	**+45.53%**	**+50.53%**
METEOR / EE-METEOR	-0.781 / 0.460	-0.667 / 0.667	-0.800 / 0.800	-0.648 / 0.035	-0.548 / 0.183	-0.632 / 0.316	-0.062 / 0.724	0.333 / 0.667	0.400 / 0.800	**+90.33%**	**+79.97%**	**+98.27%**
BERTScore / EE-BERTScore	-0.497 / 0.682	0.000 / 0.667	-0.200 / 0.800	0.567 / 0.479	0.183 / 0.183	0.316 / 0.316	0.095 / 0.895	0.333 / 1.000	0.400 / 1.000	**+63.03%**	**+44.47%**	**+53.33%**

Table 3: WMT19 system-level human correlations, for top 4 systems only. EE metrics alleviated or eliminated the phenomenon of negative correlations reported in recent literature and brought a significant improvement on correlations in **Average**.

rics: BLEU, CHRF, METEOR and BERTScore, as discussed in section 2. Enhanced versions of these metrics are denoted by EE-BLEU, EE-CHRF, EE-METEOR and EE-BERTScore, respectively, and are compared to their standard counterparts. We further compared proposed EE metrics with ESIM (Mathur et al., 2019) and YiSi-1 (Lo, 2020), since these two metrics consistently achieve remarkable performances across benchmarks of WMT19, WMT20 and WMT21. In addition, DA-BERTScore (Zhan et al., 2021a), which outperforms existing metrics in MT system evaluation owing to its unique token-level difficulty, is also involved in the comparison experiment.

Implementation Details In our implementation of EE metric, we use *fast_align*[2] (Dyer et al., 2013) to obtain aligned chunks between reference and hypothesis,.i.e., e_i, s_i in Eq. (3). For other metrics, we utilize *sacreBLEU*[3] (Post, 2018) toolkit to acquire BLEU and CHRF, and *NLTK*[4] toolkit to compute METEOR. For BERTScore[5], we use the default models except that the model for English is replaced with *deberta-xlarge-mnli* (He et al., 2021), as recommended by the authors of BERTScore.

Main Results Following the criterion of recent research (Zhan et al., 2021a; Freitag et al., 2020) as well as WMT official organization, three coefficients: Pearson's correlation r, Kendall's τ and Spearman's ρ, are used to validate system-level correlations with human DA as well as the agreement with human rankings. Values of the three coefficients range from -1 to 1, with a bigger positive value indicating a stronger positive correlation with human assessments, and a smaller negative value indicating a stronger negative correlation. Table 2 displays the main results. It can be seen that EE metrics achieve competitive correlations in the comparison. Among the enhanced metrics, EE-BERTScore further improves standard BERTScore and consistently outperforms other metrics, including DA-BERTScore and best metrics in WMT19, across different correlation measurements and translation directions. The case analysis in appendix A might help to reveal the practical meaning of the higher correlation numbers brought by EE metrics, by displaying how EE-BERTScore corrects the relative ranking of two systems given by BERTScore in En→ De. It should be noted that, even the improvement on correlations is little sometimes (e.g., r from 0.990 to 0.994 in En→ De for BERTScore), the number of corrected relative rankings between system pairs may be notable (seven more corrected cases after EE-BERTScore being applied in En→ De, similar to the one in appendix A).

The result in Table 2 shows that the four EE metrics bring average improvements of 1.65%, 4.96% and 3.18% on r, τ and ρ, respectively, compared with corresponding standard metrics across the five datasets. Despite divergent underlying mechanisms, all four backbone metrics experienced enhancement on correlations averaged across five translation tracks, which proves the universal feasibility of the proposed EE approach. The sentence-level difficulty introduced in the EE metric works as an extra dimension in system-level score aggregation, which, by assigning larger weights to high

[2]https://github.com/clab/fast_align
[3]https://github.com/mjpost/sacrebleu
[4]https://www.nltk.org/api/nltk.html
[5]https://github.com/Tiiiger/bert_score

entropy hypotheses, encourages systems that handle difficult translations well. This strategy, as well as the computation of entropy, is independent of particular MT metrics. Thus, the proposed method is compatible with a wide range of MT metrics.

Effect of Top-K Systems As reported in Ma et al. (2019), Thompson and Post (2020) and Mathur et al. (2020), in the circumstances where only top systems are preserved, most existing metrics suffer a drastic drop on correlations with human evaluations. This phenomenon is extremely notable in WMT19 En→De, De→En and Zh→En for top 4 systems, where metrics exhibit zero or even strong negative correlations with human assessments. Current research attributes this to unstable noises or outlier systems, while we found the proposed EE method helpful to alleviate the degradation of correlations owing to the extra sentence-level difficulty. In extreme competitive situations, all systems involved provide nearly perfect translations for most of the easy samples, while the high-entropy hypotheses, due to the fluctuation in translation qualities, tend to be key for humans to rank those top systems. In such a scenario, simple samples might even be harmful noises to the automatic evaluation, causing the failure of distinguishing top systems using existing metrics. In contrast, EE metrics focus on high-entropy parts in the evaluation set. Thus, as shown in Table 3, EE metrics avoid the negative correlations phenomenon (e.g., in En→De, r from -0.497 to 0.682 for BERTScore, ρ from -0.800 to 0.800 for METEOR) or even achieve perfect correlations with human rankings (e.g., in Zh→En, τ from 0.333 to 1.000, ρ from 0.400 to 1.000 for BERTScore). Averagely speaking, for top 4 systems, substantial improvements can be expected after proposed enhancement being applied.

Fig. 3 shows the process of degradation on correlations when low-performance systems are gradually removed. It can be seen that existing metrics fail to correlate with human judgments when K is smaller than 10, and start to exhibit negative correlation when K is smaller than or equal to 6. In contrast, EE-BERTScore only suffers minor drop on correlation and keeps effective with the decrease of K. The effectiveness of EE metrics further indicates the key role high-entropy samples play in an evaluation set.

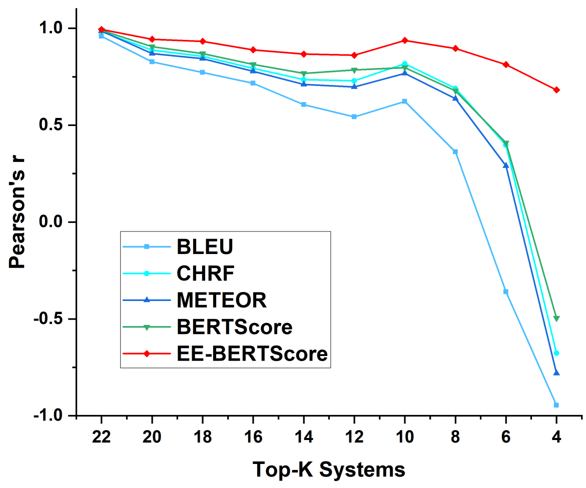

Figure 3: Effect on Pearson's correlation when only top-K systems are included in the En→De evaluation. EE-BERTScore keeps a high correlation with human judgments with the elimination of inferior systems.

5 Discussion

5.1 Estimation of Difficulty Threshold h

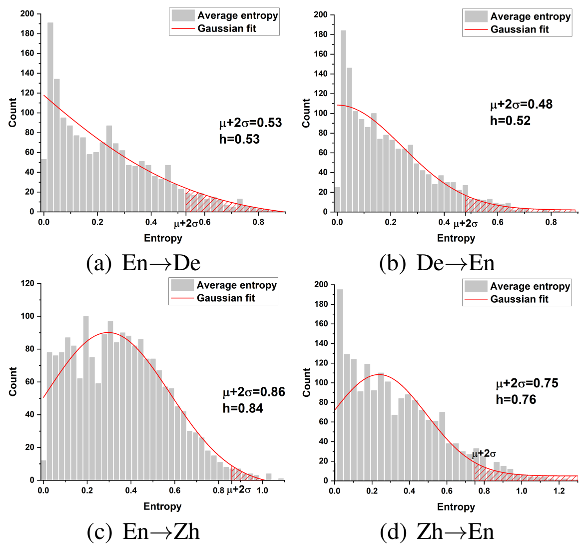

Figure 4: Distributions of mean entropy averaged across systems, i.e., $\overline{H}_{\hat{X}_{s_i}}$, extracted from (a) En→De, (b) De→En, (c) En→Zh and (d) Zh→En, fitted according to $\mathcal{N}(\mu, \sigma)$. The areas in shadow are two standard deviations away from mean values.

The parameter h functions as the threshold entropy value for a hypothesis to be classified as difficult in an evaluation set. From Eq. (5), h is estimated by examining those samples whose average translation entropy is significantly higher than others. $\overline{H}_{\hat{X}_{s_i}}$, the average entropy of sample X_{s_i}, is calculated by

$$\overline{H}_{\hat{X}_{s_i}} = \frac{1}{P}(H(X_{h_i}^{S_1}) + H(X_{h_i}^{S_2}) + ... + H(X_{h_i}^{S_P})), \quad (8)$$

where $\forall p \in [1, P], X_{h_i}^{S_p} \in \hat{X}_{s_i}$. Since $X_{h_i}^{S_p}$, the translation hypothesis of the ith source sentence produced by system p, is modeled as a random variable in Eq. (2), by central limit theorem, the distribution of $\overline{H}_{\hat{X}_{s_i}}$ can be estimated according to $\mathcal{N}(\mu, \sigma)$, assuming that P, the number of candidate systems, is large enough and $X_{h_i}^{S_1}...X_{h_i}^{S_P}$ in the translation of a certain language pair is i.i.d. Let α in Eq. (5) be 0.05. Then according to three-sigma rule of normal distribution, the two standard deviations serve as a borderline separating easy and difficult translations, with the difficult samples (around 5%) possessing significantly higher entropy. So h is estimated by

$$h = \mu + 2\sigma \tag{9}$$

Empirically obtained h is in accordance with Eq. (9), as shown in Fig. 4. We search for optimal h within a range from 0 to 1 for every language pair. For the high-resource language pairs (En$\leftrightarrow$De, En$\leftrightarrow$Zh), the group of candidate systems is relatively large, and thus $\mu + 2\sigma$ provides a good estimation of h, with an average error of only 0.018 on the four evaluation sets.

5.2 Ablation Study

Table 4 shows the result of ablation experiments conducted in order to acquire a better understanding of mechanisms of the proposed EE metric.

Approach	h	w	r	τ	ρ
BERTScore	-	-	0.990	0.807	0.931
EE-BERTScore	0.53	0.35	**0.994**	**0.859**	**0.952**
Different Thresholds					
$\boldsymbol{h = \mu + 2.5\sigma}$	0.83	0.35	0.929	0.477	0.630
$\boldsymbol{h = \mu + 1.5\sigma}$	0.23	0.35	**0.991**	**0.816**	**0.949**
Group Remove					
Only easy	0.53	1.00	0.988	0.781	0.920
Only difficult	0.53	0.00	0.990	**0.833**	**0.939**
Module Ablation					
w/o entropy	-	-	0.984	0.721	0.870
w/o difficulty	-	-	0.437	0.252	0.366

Table 4: Ablation experiment of EE-BERTScore conducted on WMT19 En$\rightarrow$De evaluation. Values in bold indicate better correlations compared to standard BERTScore.

Different Thresholds A higher threshold means fewer difficult hypotheses. When h is 2.5-σ away from mean, only most difficult samples (around 1.24%) are weighted. Since extreme high entropy is often caused by noises in references or miscalculated alignments in hypotheses, these samples cannot reflect performance of systems and thus cause a drop in agreement with human rankings. Reducing the threshold, on the other hand, amplifies contributions of some less representative segments without damaging the core difficult group and results in a minor improvement on correlations.

Group Remove By setting w to 1 or 0, difficult or easy hypotheses are zero weighted, and thus we can examine the standalone role of each group. As shown in Table 4, **completely removing the low-entropy hypotheses still leads to an improvement on correlations as compared to the standard metrics**. While this result further supports our intuition in this paper that the portion of high-entropy samples might be enough to determine the performance of MT systems, it is interesting to explore the possibility of distillation of an MT evaluation set to enhance its ability to distinguish candidates in the future.

Module Ablation Instead of calculating the entropy, we randomly divide easy and difficult groups while maintaining the original group sizes (repeated 1000 times). For the removal of difficulty, we directly compute the correlations between human ratings and average entropy of a system. The result indicates that the effectiveness of the proposed EE method relies on both entropy and sentence-level difficulty.

5.3 Stability Across MT Systems

Compared with standard reference-based metrics, which compute the score of an MT system utilizing only its hypotheses and the references, EE metrics introduce additional information of other participated systems in the computation of system-level scores, i.e., the score assigned to a certain MT system may vary with its competitors. To better understand the impact caused by the difference and possible limitations of EE metrics, we investigated the stability of EE metrics across MT systems by applying EE metrics on a series of random subsets of systems. Specifically, we randomly choose n systems (n=4,6,8,10) in En$\rightarrow$ De (22 systems) and test the correlations with human scores for all four metrics (standard and EE versions). For each n, we repeat 100 times, i.e., 100 random combinations of n systems. The results in Table 5 show that EE Metrics steadily outperform standard metrics, with average improvements of 6.90%, 8.25%,

Metric	Random 4			Random 6			Random 8			Random 10		
	r	τ	ρ	r	τ	ρ	r	τ	ρ	r	τ	ρ
BLEU	0.883	0.794	0.855	0.921	0.763	0.861	0.912	0.744	0.865	0.928	0.758	0.880
CHRF	0.902	0.744	0.819	0.945	0.780	0.879	0.944	0.789	0.895	0.959	0.784	0.898
METEOR	0.904	0.777	0.848	0.929	0.760	0.865	0.945	0.768	0.884	0.944	0.765	0.893
BERTScore	0.929	0.839	0.886	0.943	0.815	0.901	0.957	0.830	0.916	0.957	0.814	0.914
EE-BLEU	0.878	0.752	0.813	0.935	0.769	0.868	0.942	0.761	0.873	0.952	0.782	0.897
EE-CHRF	0.934	0.820	0.877	0.959	0.780	0.894	0.958	0.791	0.894	0.961	0.793	0.906
EE-METEOR	**0.945**	0.814	0.873	0.950	0.809	0.896	0.957	0.803	0.906	0.957	0.805	0.912
EE-BERTScore	**0.945**	**0.886**	**0.921**	**0.969**	**0.892**	**0.941**	**0.966**	**0.855**	**0.926**	**0.977**	**0.870**	**0.943**

Table 5: Performances of MT metrics when only **Random n** systems are involved from 22 systems in En→ De translation task. For each n, the correlations are averaged across 100 random combinations of systems.

4.59% and 6.57% on correlations, for n=4, 6, 8, 10, respectively.

6 Conclusion and Future Work

In this paper, we find that the high-entropy hypotheses, though holding only a minor portion in an evaluation set, play a significant role in terms of correlations with human judgments in MT evaluation. By rebalancing the weights between low-entropy and high-entropy hypotheses, an entropy enhancing approach for MT metrics is proposed. Experimental results on five sub-tracks in WMT19 metric tasks show that our proposed approach successfully enhances the performance of popular MT metrics and achieves remarkable correlations with human assessments, especially in the evaluation of competitive systems. Our analysis introduces the concept of sentence-level difficulty into MT evaluation and reveals the importance of difficult samples in system-level evaluations.

There are several directions for future exploration. First, entropy-based difficulty can work as a measurement to the quality of an MT evaluation set. If an evaluation set contains more high-entropy samples, its ability to rank systems is better. Second, using entropy, we can dig the hard samples out of an evaluation set and, by filtering easy samples, we can make a distillation of evaluation set. Third, there is still room for optimization in calculation of entropy and difficulty weights.

References

Rafael E. Banchs, Luis F. D'Haro, and Haizhou Li. 2015. Adequacy–fluency metrics: Evaluating mt in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):472–482.

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

Julian Chow, Lucia Specia, and Pranava Madhyastha. 2019. WMDO: Fluency-based word mover's distance for machine translation evaluation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 494–500, Florence, Italy. Association for Computational Linguistics.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, page 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

C. Dyer, V. Chahuneau, and N. A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. *proc naacl*.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2006. CDER: Efficient MT evaluation using block movements. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 241–248, Trento, Italy. Association for Computational Linguistics.

Xuebo Liu, Houtim Lai, Derek F Wong, and Lidia S Chao. 2020. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 427–436.

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.

Chi-kiu Lo. 2020. Extended study on using pretrained language models and yisi-1 for machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 895–902.

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.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2019. Putting evaluation in context: Contextual embeddings improve machine translation evaluation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2799–2808.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

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.

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.

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

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

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.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

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.

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.

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.

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.

Runzhe Zhan, Xuebo Liu, Derek F Wong, and Lidia S Chao. 2021b. Meta-curriculum learning for domain adaptation in neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14310–14318.

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*.

Yang Zhao, Jiajun Zhang, Chengqing Zong, Zhongjun He, and Hua Wu. 2019. Addressing the under-translation problem from the entropy perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 451–458.

A Case Study

The two cases in Table 6 illustrate how EE-BERTScore enhances the performance of BERTScore via the discussed strategy. The system-level score of MSRA's translation system, given by BERTScore, is higher than that of Facebook's, leading to a misalignment with human rankings (Facebook ranks the 1st in En→De while MSRA ranks the 4th). In contrast, EE-BERTScore successfully recognizes Facebook as the superior system. From Table 6, Facebook outperforms MSRA in difficult translations (Case 1), despite defeated in easier sentences (Case 2). In BERTScore, the difference of segments are ignored and all segment-level scores are of the same contribution to the final system score. As a result, the final score of Facebook is slightly lower than MSRA. In human evaluation, ratings for simple hypotheses produced by different systems tend to similar, because these hypotheses are already in good alignment with the reference. While scores of the difficult ones, implying a challenging segment in source language, often separate top systems from inferior candidates. Utilizing this strategy, EE-BERTScore amplified the contribution of difficult segments in case 1 for both systems (0.039%→0.276%, 0.042%→0.311%), while reduces the contribution of simpler hypotheses (0.037%→0.015%, 0.034%→0.013%). Consequently, Facebook exceeded MSRA owing to its advantages in difficult hypotheses.

As discussed in section 3.2, in the proposed method, determination of sentence-level difficulty relies on entropy values. In Table 6, entropy values of hypotheses in case 1 are higher than h, the threshold determined by Eq. (5), while the easy hypotheses in case 2 hold smaller values of entropy. The reason is that hypotheses in case 2 are divided into smaller groups of aligned chunks, and the lengths of chunks are more evenly distributed, as highlighted by the colored boxes, implying a less disordered distribution of hypothesis and lower entropy of translation.

B Estimation of Coefficient w

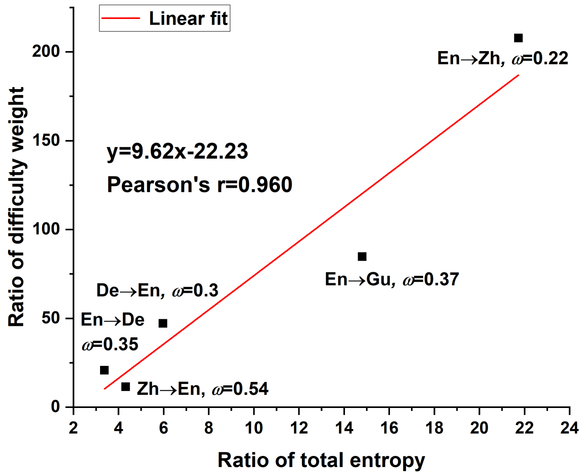

Figure 5: An empirical fit of Eq. (10). The x-axis, **Ratio of total entropy**, represents the right side of Eq. (10), and y-axis denotes left side of Eq. (10). Data points are computed based on the five WMT19 evaluation sets and corresponding empirically obtained w.

The determination of sentence-level difficulty weight, i.e., W in Eq. (4), relies on h and w. In section 5.1, based on definitions in Eq. (5), we pre-

	BERTScore		EE-BERTScore		Sentence	Entropy
	Seg. / Sys.	Contrib.	Seg. / Sys.	Contrib.		
Case 1: Difficult sentence contribute more in calculation of EE-BERTScore						
Src	-	-	-	-	Likening the suit to "extortion," Plasco said his wife was just two months off having a baby and was in a "very difficult situation."	-
Ref	-	-	-	-	Plasco sagte, dass seine Frau im siebten Monat schwanger und nicht in bester Verfassung gewesen sei, und bezeichnete die Klage als „Erpressung“.	-
MSRA	0.648 / **0.830**	0.039%	0.648 / 0.799	0.276%	Plasco verglich den Anzug mit „Erpressung“ und sagte, seine Frau sei nur zwei Monate von einem Baby entfernt und befinde sich in einer „sehr schwierigen Situation“.	0.663
Facebook	0.689 / 0.828	0.042%	0.689 / **0.801**	0.311%	Plasco verglich die Klage mit „Erpressung“ und sagte, seine Frau habe gerade zwei Monate kein Baby bekommen und befinde sich in einer „sehr schwierigen Situation“.	0.642
Case 2: Easy sentence contribute less in calculation of EE-BERTScore						
Src	-	-	-	-	When that momentum gets going one way, it puts a lot of pressure on those middle matches.	-
Ref	-	-	-	-	Wenn sich erstmal eine Eigendynamik entwickelt hat, übt das großen Druck auf die mittleren Matches aus.	-
MSRA	0.609 / **0.830**	0.037%	0.609 / 0.799	0.015%	Wenn diese Dynamik in eine Richtung geht, übt sie viel Druck auf diese mittleren Spiele aus.	0.459
Facebook	0.555 / 0.828	0.034%	0.555 / **0.801**	0.013%	Wenn dieses Momentum in eine Richtung geht, setzt es diese mittleren Spiele stark unter Druck.	0.226

Table 6: Examples from the En→ De evaluation, where EE-BERTScore corrects the ranking of two systems given by BERTScore. Seg. and Sys. denotes segment-level and system-level scores given by metric, respectively, and Contrib. denotes contribution of the particular segment to final system score(e.g. $0.039\% = 0.648 \div 1997 \div 0.830, 0.311\% = 0.689 \times 0.65 \div 180 \div 0.801$). The difficulty level of cases are determined by their entropy value. Chunks indicate the alignments with reference.

sented an estimation of optimal h. Now, w, the balancing coefficient which is introduced to alleviate the distortion caused by unbalanced size between the difficult group and easy group, theoretically satisfies the following equation:

$$\frac{(1-w)(L-|D_s|)}{w|D_s|} \propto \frac{\sum_{t=1, X_{s_t} \notin D_s}^{L} \overline{H}_{\hat{X}_{s_t}}}{\sum_{k=1, X_{s_k} \in D_s}^{L} \overline{H}_{\hat{X}_{s_k}}} \tag{10}$$

Eq. (10) guarantees that the weights W assigned to difficult group and easy group are determined by the ratio of average entropy in two groups. From Eq. (10), difficulty weight W on a particular evaluation set is fully determined by distribution of average entropy within a given dataset, via different balancing coefficients w. When the total entropy of difficult samples in an evaluation set decreases, which means the translations in this evaluation set are easier, the weights assigned on difficult samples should also be higher to better distinguish difficult hypotheses from easy ones. In experiment, we search for optimal w within a range from 0 to 1 for every language pair. The empirically obtained optimal w is highly related to the statistics described in Eq. (10) with $|r| = 0.960$, as shown in Fig. 5. Linear fit based on the five WMT19 evaluation sets provides an empirical estimation of w:

$$w = \frac{R_{\overline{N}}}{9.62 R_{\overline{H}} + R_{\overline{N}} - 22.23} \tag{11}$$

where $R_{\overline{H}} = \frac{\sum\{\overline{H}_{\hat{X}_{s_t}} \mid \forall t \in [1,L], X_{s_t} \notin D_s\}}{\sum\{\overline{H}_{\hat{X}_{s_k}} \mid \forall k \in [1,L], X_{s_k} \in D_s\}}$, $R_{\overline{N}} = \frac{L-|D_s|}{|D_s|}$, are defined in Eq. (10) and fully determined by distribution of translation entropy within an evaluation set.

C Parameters

Parameters	En→ De	De→ En	En→ Zh	Zh→ En	En→ Gu
h	0.53	0.52	0.84	0.76	0.72
w	0.35	0.30	0.22	0.54	0.37

Table 7: Parameters used in our experiment. All experimentally acquired parameters are in accordance with our theoretical analysis.

D Additional Experimental Results

Metric	En→ De			Zh→ En		
	r	τ	ρ	r	τ	ρ
BLEU	0.831	0.714	0.821	0.360	0.357	0.571
CHRF	0.917	0.810	0.893	0.425	0.357	0.524
METEOR	0.854	0.619	0.714	0.678	0.643	0.738
BERTScore	0.754	0.429	0.536	0.742	0.643	0.810
EE-BLEU	0.810	0.714	0.821	0.322	0.214	0.405
EE-CHRF	0.890	0.810	0.893	0.510	0.357	0.524
EE-METEOR	0.805	0.619	0.714	0.770	0.786	0.857
EE-BERTScore	0.724	0.429	0.536	0.895	0.714	0.833

Table 8: Performances of EE Metrics on WMT 2020 news test (without human), using human MQM scores as the ground truth. Parameters h and w are computed according to Eq. 9 and Eq. 11. The result shows an average of 2.67 % improvements on correlations with human MQM scores after the enhancement on the standard metrics being applied.

Metric	En→ De			Zh→ En			En→ Ru		
	r	τ	ρ	r	τ	ρ	r	τ	ρ
BLEU	0.918	0.897	0.967	0.549	0.282	0.429	0.576	0.385	0.521
CHRF	0.813	0.692	0.868	0.366	0.154	0.297	0.768	0.451	0.653
METEOR	0.813	0.718	0.885	0.432	0.282	0.385	0.772	0.495	0.670
BERTScore	0.911	0.795	0.945	0.577	0.308	0.484	0.776	0.538	0.692
EE-BLEU	0.910	0.821	0.934	0.528	0.333	0.484	0.720	0.451	0.587
EE-CHRF	0.764	0.692	0.857	0.361	0.231	0.313	0.725	0.560	0.741
EE-METEOR	0.869	0.718	0.874	0.416	0.231	0.308	0.784	0.582	0.736
EE-BERTScore	0.876	0.846	0.945	0.630	0.487	0.626	0.655	0.473	0.644

Table 9: Performances of EE Metrics on WMT 2021 news test (without human), using human MQM scores as the ground truth and ref A as the reference. Parameters h and w are computed according to Eq. 9 and Eq. 11. The result shows an average of 4.48 % improvements on correlations with human MQM scores after the enhancement on the standard metrics being applied.