KG-BERTScore: Incorporating Knowledge Graph into BERTScore for Reference-Free Machine Translation Evaluation

Zhanglin Wu, Min Zhang, Ming Zhu, Yinglu Li, Ting Zhu, Hao Yang*, Song Peng, Ying Qin
{wuzhanglin2,zhangmin186,zhuming47,liyinglu,zhuting20,yanghao30,pengsong2,qinying}@huawei.com
Huawei Technologies Co., Ltd.
Beijing, China

ABSTRACT

BERTScore is an effective and robust automatic metric for reference-based machine translation evaluation. In this paper, we incorporate multilingual knowledge graph into BERTScore and propose a metric named KG-BERTScore, which linearly combines the results of BERTScore and bilingual named entity matching for reference-free machine translation evaluation. From the experimental results on WMT19 QE as a metric without references shared tasks, our metric KG-BERTScore gets higher overall correlation with human judgements than the current state-of-the-art metrics for reference-free machine translation evaluation.[1] Moreover, the pre-trained multilingual model used by KG-BERTScore and the parameter for linear combination are also studied in this paper.

KEYWORDS

machine translation evaluation; multilingual knowledge graph; BERTScore; KG-BERTScore; pre-trained multilingual model

ACM Reference Format:
Zhanglin Wu, Min Zhang, Ming Zhu, Yinglu Li, Ting Zhu, Hao Yang*, Song Peng, Ying Qin. 2022. KG-BERTScore: Incorporating Knowledge Graph into BERTScore for Reference-Free Machine Translation Evaluation. In *Proceedings of October 27–29, 2022 (International Joint Conference on Knowledge Graphs).* Hangzhou, China, 5 pages.

1 INTRODUCTION

Machine translation (MT) evaluation is an important research topic in natural language processing, and its development plays a crucial role in the progress of machine translation. Although Human judgement is an ideal MT evaluation metric, automatic MT evaluation metrics are applied in most cases due to the former's long evaluation cycle and high labor consumption. With the continuous deepening of research, automatic MT evaluation metrics have been divided into reference-based MT evaluation metrics and reference-free MT evaluation metrics[3].

Reference-based MT evaluation metrics are mainly separated into three categories: n-gram similarity based-metrics, editing distance-based metrics and word embedding-based metrics. N-gram similarity-based metrics include BLEU[18], chrF[20] and METEOR[1], etc. Among these metrics, BLEU measures the correspondence of n-grams between machine translation and reference, chrF uses character n-gram instead of word n-gram, and METEOR takes into account the form of words, expands thesaurus with knowledge sources such as WordNet[16]. Editing distance-based metrics such as TER[23], WER[17] and PER[17] evaluate the quality based on the minimum number of edits required to convert a machine translation into a reference, among which the difference lies in the definition of "error" and the type of editing action. The above two types of reference-based metrics can only provide rule-based metrics while word embedding-based metrics can also take into account the intrinsic meaning of words. Word embedding-based metrics are divided into semantic information-based metrics and end-to-end evaluation metrics. The semantic information-based metrics use pre-training model such as word2vector[15] or BERT[9] to perform lexical level analysis and alignment of machine translation and reference to calculate the semantic similarity, whose implementations include MEANT 2.0[12], YiSI[13], BLEURT[22] and BERTScore[29], etc.[2] The end-to-end evaluation metrics usually adopt predictor-estimator[11] architecture and use multilingual pre-training model such as XLM-R[5] as predictor for encoding and estimator for scoring. These metrics such as COMET[26], rely on human scoring data to train the model.

Reference-free MT evaluation metrics are more challenging and promising compared with reference-based ones. Early reference-free MT evaluation metrics like QuEST[25] and QuEST++[24] are heavily dependent on linguistic processing and feature engineering to train traditional machine-learning algorithms like support vector regression and randomised decision trees[25]. Such metrics are called artificial feature based reference-free MT evaluation metrics. In effect, these metrics are generally considered to be inferior to neural network based reference-free MT evaluation metrics. Neural network based reference-free MT evaluation metrics refer to using neural network for end-to-end modeling, automatically extracting features, and then evaluating the quality of machine translation. Early neural network based reference-free MT evaluation metrics include POSTECH[11] and deepquest[8], which require pre-training with large-scale parallel corpus. In the OpenKiwi[10] framework, there is an improved metric based on the pre-trained language model, which avoids relying on large-scale parallel corpus and requires only a small amount of human scoring data to fine-tune neural network. On this basis, TransQuest[21] selects XLM-R as the pre-training language model, and simplifies the structure of neural network, which improves the computational efficiency and evaluation accuracy.

[1] https://www.statmt.org/wmt19/qe-task.html

[2] https://github.com/Tiiiger/bert_score

International Joint Conference on Knowledge Graphs , Hangzhou, China,

Table 1: A KG-BERTScore calculation example for en-zh. F_{BERT} is calculated by xlm-roberta-base, and α is set to 0.5.

source sentence	Respiratory irritation was not reported in Northwest Florida over the past week.
corresponding entity IDs	/m/0hl_6 **/m/02xry** /m/083sl
machine translation	本周，佛罗里达西北部没有消化道刺激的报告。
corresponding entity IDs	/m/05qv5f **/m/02xry** /m/0j49l /m/0chln1
F_{BERT}	0.857
F_{KG}	$\frac{1}{3} = 0.333$
$F_{KG-BERT}$	$0.333 \times 0.5 + 0.857 \times (1-0.5) = 0.595$

MT evaluation metrics focus on how to improve the correlation between evaluation results with human judgements. BERTScore is a typical reference-based MT evaluation metric, which correlates better with human judgements. It can also be used as a reference-free MT evaluation metric by embedding words using pre-trained multilingual model, but it correlates badly with human judgement. Therefore, we incorporate multilingual knowledge graph[4] into BERTScore for reference-free MT evaluation, and propose a reference-free metric KG-BERTScore. Our metric uses multilingual knowledge graph and pre-trained multilingual model instead of fine-tuning on parallel corpora and human scoring data, which can accurately evaluate machine translations.

To summarize, our work includes the following contributions:

- To the best of our knowledge, we are the first to focus on how to combine multilingual knowledge graph with pre-trained multilingual model for reference-free MT evaluation.
- We propose an unsupervised metric KG-BERTScore, which incorporates multilingual knowledge graph into BERTScore for reference-free MT evaluation.
- We show that KG-BERTScore correlates better with human judgment on the WMT19 QE as a metric without references shared task[6] than the current state-of-the-art reference-free MT evaluation metrics.

2 METHODS

2.1 BERTScore

BERTScore is a reference-based MT evaluation metric. We try to use it for reference-free MT evaluation and the specific steps to generate a system-level score can be described as follows:

First, the word embedding is generated by pre-trained multilingual model, and then the cosine similarity $x_i^T \hat{x}_j$ of each word x_i in source text and each word $\hat{x}_j$ in machine translation is calculated. We use greedy matching to maximize the cosine similarity score, where each word matches the most similar word in another sentence.

Then, we calculate F_{BERT} Score for each machine translation sentence as follows:

$$R = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j \tag{1}$$

$$P = \frac{1}{|\hat{x}|} \sum_{\hat{x}_i \in \hat{x}} \max_{x_j \in x} \hat{x}_i^T x_j \tag{2}$$

$$F_{BERT} = 2\frac{P \cdot R}{P + R} \tag{3}$$

Finally, we average F_{BERT} Score of all machine translation sentences to obtain a system-level score.

2.2 KG-BERTScore

We put forward a reference-free KG-BERTScore MT evaluation metric, which incorporates multilingual knowledge graph into BERTScore for reference-free MT evaluation. The evaluation process is shown in Algorithm 1:

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 $\{s_k, t_k\} \in \{S, T\}$ do
    // $x_i$, $x_j$, $\hat{x}_i$, $\hat{x}_j$ is the word embedding.
    $R_k = \frac{1}{|s_k|} \sum_{x_i \in s_k} \max_{\hat{x}_j \in t_k} x_i^T \hat{x}_j$
    $P_k = \frac{1}{|t_k|} \sum_{\hat{x}_i \in t_k} \max_{x_j \in s_k} \hat{x}_i^T x_j$
    $F_{BERT_k} = 2\frac{P_k \cdot R_k}{P_k + R_k}$
    // $entities(s_k)$, $entities(t_k)$ is the number of
       entities.
    if $entities(s_k) \neq 0$ then
        $F_{KG_k} = \frac{matches(entities(s_k), entities(t_k))}{entities(s_k)}$
    else
        $F_{KG_k} = 1$
    end
    // $\alpha$ is an adjustable hyperparameter.
    $F_{KG-BERT_k} = \alpha \cdot F_{KG_k} + (1 - \alpha) \cdot F_{BERT_k}$
end
$F = \frac{\sum_{k=1}^{n} F_{KG-BERT_k}}{n}$
```

Firstly, we employ reference-free BERTScore metric to calculate F_{BERT} score of each machine translation sentence.

Secondly, we annotate the named entities and the corresponding entity IDs in the sentences and calculate the entity matching scores. We can utilize named entity recognition model such as W-NER[27] to identify named entities, and entity links[14] to retrieve their 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. Specifically, for source sentence s and machine translation sentence t, the F_{KG} score is calculated as follows:

$$F_{KG} = \frac{matches\left(entities\left(s\right), entities\left(t\right)\right)}{entities\left(s\right)}. \tag{4}$$

Then, the above two scores are combined to obtain $F_{KG-BERT}$ score as the final evaluation result of machine translation sentence.

$$F_{KG-BERT} = \alpha \cdot F_{KG} + (1 - \alpha) \cdot F_{BERT} \tag{5}$$

Finally, we average $F_{KG-BERT}$ score of all machine translation sentences to obtain a system-level score.

Table 1 shows a KG-BERTScore calculation example for en-zh language pair. In subsequent experiments, if α parameters in the formula are not described, the default value is 0.5, and if there is no entity in the source, F_{KG} score is 1.

3 EXPERIMENTS

We conduct a comparative experiment on WMT19 QE as a metric without references shared task to test the effectiveness of our reference-free MT evaluation metric. We use xlm-roberta-base as the default pre-trained multilingual model and the ninth layer of the model for word embedding to calculate F_{BERT} scores.[3] As for the calculation of F_{KG} score, following the method of Zorik et al[7], we use Google Knowledge Graph Search API to annotate named entities and their entity IDs in sentences.[4] The annotated data can be downloaded from http://storage.googleapis.com/gresearch/kobe/data/annotations.zip.

3.1 Datasets

We collect the source sentences and system translation sentences from WMT19 news translation shared task, which contains 233 translation systems across 18 language pairs.[5] Each language pair has approximately 1,000-2,000 source sentences.

3.2 Baselines

For each language pair, we apply reference-free BERTScore and KG-BERTScore to score translation systems in WMT19 news translation shared task[2] respectively. We then measure the Pearson correlation of these two scores with human judgements. Finally, we compare a series of reference-free MT evaluation metrics: ibm1-morpheme and ibm1-pos4gram[19], LASER[28], LogProb[28], YiSi-2 and YiSi-2-srl[13], and a reference-based MT evaluation metric: BLEU.

3.3 Results

The results for language pairs into English are available in Table 2. Reference-free KG-BERTScore outperforms all other reference-free MT evaluation metrics for de-en, gu-en, kk-en, lt-en and ru-en. The average Pearson correlation of reference-free KG-BERTScore on all language pairs into English is 0.830, only 0.077 lower than that of BLEU. Table 3 describes the results for language pairs translated from English, reference-free KG-BERTScore outperforms for en-kk and en-zh. Besides, the results for language pairs not involving English are available in Table 4. In this case, reference-free KG-BERTScore outperforms for de-cs with Pearson correlation of 0.959. In conclusion, reference-free KG-BERTScore has a higher overall pearson correlation with human judgements than reference-free BERTScore metric and the other metrics we know for reference-free MT evaluation.

Table 2: System-level pearson correlation with human judgements for language pairs into English from the WMT19 QE as a metric without references shared task.

src-mt	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
LASER	0.247	-	-	-	-	-0.310	-	-
LogProb	-0.474	-	-	-	-	-0.488	-	-
ibm1-morpheme	0.345	0.740	-	-	0.487	-	-	-
ibm1-pos4gram	0.339	-	-	-	-	-	-	-
UNI	0.846	**0.930**	-	-	-	0.805	-	-
UNI+	0.850	0.924	-	-	-	0.808	-	-
YiSi-2	0.796	0.642	-0.566	-0.324	0.442	-0.339	0.940	0.227
YiSi-2 srl	0.804	-	-	-	-	-	**0.947**	-
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 3: System-level pearson correlation with human judgements for language pairs from English from the WMT19 QE as a metric without references shared task.

Metric	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh	mean
BLEU	0.897	0.921	0.969	0.737	0.852	0.989	0.986	0.901	0.907
LASER	-	0.871	-	-	-	-	-0.823	-	-
LogProb	-	-0.569	-	-	-	-	-0.661	-	-
ibm1-morpheme	**0.871**	0.870	0.084	-	-	**0.810**	-	-	-
ibm1-pos4gram	-	0.393	-	-	-	-	-	-	-
UNI	0.028	0.841	**0.907**	-	-	-	**0.919**	-	-
UNI+	-	-	-	-	-	-	0.918	-	-
USFD	-	-0.224	-	-	-	-	0.857	-	-
USFD-TL	-	-0.091	-	-	-	-	0.771	-	-
YiSi-2	0.324	0.924	0.696	0.314	0.339	0.055	-0.766	-0.097	0.224
YiSi-2 srl	-	**0.936**	-	-	-	-	-	-0.118	-
BERTScore	0.035	0.893	0.765	**0.549**	0.650	-0.084	-0.779	-0.127	0.238
KG-BERTScore	0.364	0.897	0.595	-0.197	**0.839**	-0.081	0.638	**0.077**	**0.392**

Table 4: System-level pearson correlation with human judgements for language pairs excluding English from the WMT19 QE as a metric without references shared task.

Metric	de-cs	de-fr	fr-de	mean
BLEU	0.941	0.891	0.864	0.899
ibm1-morpheme	0.355	-0.509	-0.625	-0.260
ibm1-pos4gram	-	0.085	**-0.478**	-
YiSi-2	0.606	**0.721**	-0.530	0.266
BERTScore	0.572	0.692	-0.746	0.173
KG-BERTScore	**0.959**	0.556	-0.713	**0.267**

In addition, we also notice that KG-BERTScore does not perform very well on language pairs such as en-gu, en-lt and fr-de,

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

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

[5] https://www.statmt.org/wmt19/translation-task.html

which is due to the insufficient embedding conversion ability of pre-trained multilingual model and the weak named entities coverage of multilingual knowledge graph.

4 ANALYSIS

Summarizing the above findings, reference-free KG-BERTScore obtains the best results on 8 out of 18 language pairs. This means that incorporating multilingual knowledge graph into BERTScore is a promising path towards reference-free MT evaluation. In this section, we analyze the factors that affect the effectiveness of the reference-free KG-BERTScore metric.

4.1 Impact of Different Pre-training Multilingual Models

To measure the impact of different pre-trained multilingual models on reference-free BERTScore and KG-BERTScore, we select several commonly used pre-trained l models: bert-base-multilingual-cased,[6] xlm-roberta-base, xlm-roberta-large.[7] Based on these l models, reference-free BERTScore and KG-BERTScore are employed to evaluate the language pairs into English from WMT19 QE as a metric without references shared task. The average pearson correlation between the evaluation results of all language pairs and human judgments is shown as Figure 1.

Experimental results show that while the pre-trained multilingual model performs better on reference-free BERTScore metric, it also performs better on reference-free KG-BERTScore metric. Furthermore, reference-free KG-BERTScore metric is consistently more accurate than reference-free BERTScore metric under the same pre-trained multilingual model.

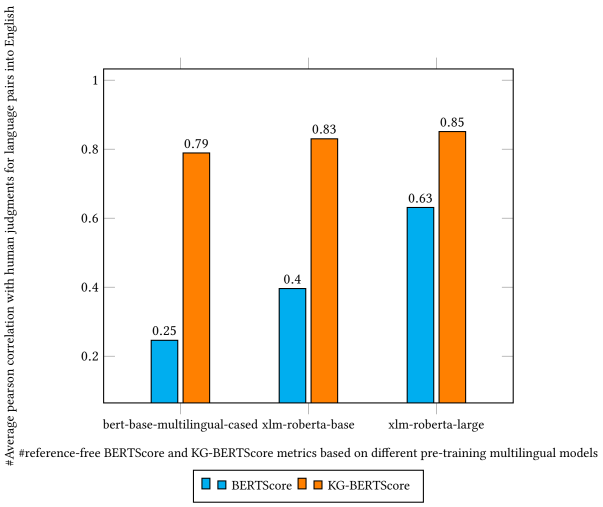

Figure 1: Average system-level pearson correlation with human judgements of reference-free BERTScore and KG-BERTScore metrics based on different pre-training multilingual models for language pairs into English.

4.2 Impact of Different weights in reference-free KG-BERTScore

To explore the impact of different weights in reference-free KG-BERTScore metric, we apply the reference-free KG-BERTScore metrics with weights of 0.0, 0.2, 0.4, 0.5, 0.6, 0.8, 1.0 to evaluate the language pairs into English from WMT19 QE as a metric without references shared task. Table 5 indicates the pearson correlation between the evaluation results and human judgments for each language pair. The experimental results show that the combination of knowledge graph and BERTScore is better than that of only knowledge graph or BERTScore, and when the weight is 0.5, the overall evaluation accuracy is close to the best.

Table 5: System-level pearson correlation with human judgements of reference-free KG-BERTScore metrics with different weights for language pairs into English.

src-mt	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en	mean
α=0.0	**0.867**	0.656	0.663	0.885	0.642	0.898	0.912	0.789
α=0.2	0.852	0.857	0.547	0.867	0.686	0.766	0.891	0.781
α=0.4	0.861	0.774	0.739	0.926	**0.688**	0.904	0.906	0.828
α=0.5	0.862	0.733	0.764	0.936	**0.688**	0.918	0.908	**0.830**
α=0.6	0.864	0.696	0.774	0.943	**0.688**	0.924	0.910	0.828
α=0.8	0.865	0.636	**0.778**	**0.950**	**0.688**	**0.930**	**0.913**	0.823
α=1.0	0.785	**0.866**	-0.007	0.117	0.657	-0.372	0.728	0.396

5 CONCLUSION

In the paper, a reference-free KG-BERTScore metric is proposed for MT evaluation. Compared with traditional metrics, the metric is unsupervised and does not require parallel corpus and human scoring data for pre-training and fine-tuning, but only requires multilingual knowledge graph and pre-trained multilingual model. We also verify the effectiveness of KG-BERTScore on WMT19 QE as a metric without references shared task, and its experimental results show that the metric is reliable and promising.

REFERENCES

[1] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[2] Loïc Barrault, Ondřej Bojar, Marta R Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. 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)*. 1–61.

[3] Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*. 12–58.

[4] 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*. 1511–1517.

[5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard 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*. 8440–8451.

[6] https://huggingface.co/bert-base-multilingual-cased

[7] https://huggingface.co/xlm-roberta-large

[6] Erick Fonseca, Lisa Yankovskaya, André FT Martins, Mark Fishel, and Christian Federmann. 2019. Findings of the WMT 2019 shared tasks on quality estimation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*. 1–10.

[7] Zorik Gekhman, Roee Aharoni, Genady Beryozkin, Markus Freitag, and Wolfgang Macherey. 2020. KoBE: Knowledge-Based Machine Translation Evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 3200–3207.

[8] Julia Ive, Frédéric Blain, and Lucia Specia. 2018. DeepQuest: a framework for neural-based quality estimation. In *Proceedings of the 27th International Conference on Computational Linguistics*. 3146–3157.

[9] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[10] Fábio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André FT Martins. 2019. OpenKiwi: An Open Source Framework for Quality Estimation. *ACL 2019* (2019), 117.

[11] Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. In *Proceedings of the Second Conference on Machine Translation*. 562–568.

[12] Chi-kiu Lo. 2017. MEANT 2.0: Accurate semantic MT evaluation for any output language. In *Proceedings of the second conference on machine translation*. 589–597.

[13] 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)*. 507–513.

[14] Weiming Lu, Peng Wang, Xinyin Ma, Wei Xu, and Chen Chen. 2020. Enrich cross-lingual entity links for online wikis via multi-modal semantic matching. *Information Processing & Management* 57, 5 (2020), 102271.

[15] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. (2013).

[16] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[17] Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st annual meeting of the Association for Computational Linguistics*. 160–167.

[18] 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*. 311–318.

[19] Maja Popović. 2012. Morpheme-and POS-based IBM1 and language model scores for translation quality estimation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*. 133–137.

[20] Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. 392–395.

[21] Tharindu Ranasinghe, Constantin Or'Ősan, and Ruslan Mitkov. 2020. TransQuest: Translation Quality Estimation with Cross-lingual Transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*. 5070–5081.

[22] 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*. 7881–7892.

[23] Matthew Snover, Bonnie Dorr, Richard 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*. 223–231.

[24] Lucia Specia, Gustavo Paetzold, and Carolina Scarton. 2015. Multi-level translation quality prediction with quest++. In *Proceedings of ACL-IJCNLP 2015 system demonstrations*. 115–120.

[25] Lucia Specia, Kashif Shah, José GC De Souza, and Trevor Cohn. 2013. QuEst-A translation quality estimation framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 79–84.

[26] Craig Stewart, Ricardo Rei, Catarina Farinha, and Alon Lavie. 2020. COMET-Deploying a New State-of-the-art MT Evaluation Metric in Production.. In *AMTA (2)*. 78–109.

[27] Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A Unified Generative Framework for Various NER Subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5808–5822.

[28] Elizaveta Yankovskaya, Andre Tättar, and Mark Fishel. 2019. Quality estimation and translation metrics via pre-trained word and sentence embeddings. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*. 101–105.

[29] 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*.