IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 March 2023, accepted 18 March 2023, date of publication 30 March 2023, date of current version 4 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3260835

RESEARCH ARTICLE

Implicit Cross-Lingual Word Embedding Alignment for Reference-Free Machine Translation Evaluation

MIN ZHANG, (Member, IEEE), HAO YANG, (Senior Member, IEEE), YANQING ZHAO, XIAOSONG QIAO, SHIMIN TAO, SONG PENG, YING QIN, AND YANFEI JIANG
Huawei Translation Services Center, Beijing 100095, China
Corresponding author: Min Zhang (zhangmin186@huawei.com)

ABSTRACT As we know, cross-lingual word embedding alignment is critically important for reference-free machine translation evaluation, where source texts are directly compared with system translations. In this paper, it is revealed that multilingual knowledge distillation for sentence embedding alignment could achieve cross-lingual word embedding alignment implicitly. A simplified analysis is given to explain the implicit alignment reason. And according to the analysis, it could be deduced that using the last layer embeddings of the distilled student model will have the best alignment effect, which is also validated by the experimental results on the WMT19 datasets. Furthermore, with the assistant of a target-side language model, BERTScore and Word Mover's Distance using the cross-lingual word embeddings get very competitive results (4 best average scores on 3 types of language directions and ranking first among more than half of all 18 language pairs for the system-level evaluations) in the WMT19's reference-free machine translation evaluation tasks when the current state-of-the-art (SOTA) metrics are chosen for comparison.

INDEX TERMS Cross-lingual word embedding, machine translation evaluation, multilingual knowledge distillation, target-side language model, word embedding alignment.

I. INTRODUCTION

Reference texts are provided and compared with system translations in traditional machine translation (MT) evaluation methods, such as those used by the famous *n*-gram based metric BLEU [1] and recent word embedding based metrics BERTScore [2] and BLEURT [3].

However, reference sentences could only cover a tiny fraction of input source sentences, and non-professional translators can not yield high-quality human reference translations [4]. Recently, with the rapid progress of deep learning in multilingual language processing [5], [6], [7], [8], [9], [10], [11], [12], there has been a growing interest in reference-free MT evaluation, which is also referred to as ''quality estimation'' (QE) in the MT community [13]. In QE, evaluation metrics compare system translations with source sentences directly. Therefore, the alignment between source sentences and system translations is crucial for reference-free MT evaluation. To the best of our knowledge, previous works focus on the direct alignments on cross-lingual lexical, word embedding or sentence embedding levels [14], [15], [16], [17], [18], [19], [20].

In this paper, we highlight that cross-lingual word embedding alignment could be achieved implicitly by multilingual knowledge distillation (MKD) for sentence embedding alignment [21]. The reason why the alignment could be achieved implicitly is theoretically analyzed under a simplified condition. And from the analysis, it is drawn that the word embeddings in the last layer of the distilled student model have the best alignment effect. We choose 8 language pairs from WMT19 datasets to validate this, and the experimental results are in complete agreement with the theoretical analysis. Moreover, from the experimental results on all the 18 language pair datasets of WMT19, BERTScore [2] and Word Mover's Distance (WMD) [22] using the cross-lingual word embeddings are competitive metrics for both segment-level

The associate editor coordinating the review of this manuscript and approving it for publication was Jolanta Mizera-Pietraszko.

and system-level reference-free MT evaluations. With the assistant of the target-side language model [23], the two reference-free metrics show much greater competitiveness by getting the best average scores on the majority of the language directions of WMT19.

Overall, the main contributions of this paper are as follows:

- To the best of our knowledge, we are the first to highlight the cross-lingual word embedding alignment could be implicitly achieved by MKD.
- With these cross-lingual word embeddings, we design two reference-free metrics for MT evaluation under the frameworks of BERTScore and WMD.
- With the assistant of the target-side language model proposed in [23], the two metrics get very competitive results for both segment-level and system-level MT evaluations on WMT19.

A shorter conference version of this paper appeared in [24]. This initial conference paper did not provide a quantitative analysis for MKD, design reference-free metrics for MT evaluation under the framework of WMD or introduce the target-side language model to the designed metrics. This manuscript addresses these issues — it provides a quantitative analysis for MKD and designs much better reference-free metrics under the frameworks of BERTScore and WMD with the help of the target-side language model [23].

II. RELATED WORK

With the rapid progress in deep learning for machine translation [25], various methods have been proposed for reference-free MT evaluation [14], [15], [16], [17], [18], [19], [20], [26], [27], [28], [29], [30]. These methods can be divided into three main representative directions.

The first direction is lexicon-based methods that align words or named entities between source sentences and system translations. Popović et al. exploited a bag-of-word translation model for quality estimation, which sums over the likelihoods of aligned word pairs between source and translation texts [14]. Specia et al. used language-agnostic linguistic features extracted from source texts and system translations to estimate quality [15]. Gekhman et al. proposed a simple and effective Knowledge-Based Evaluation (KoBE) method by measuring the recall of entities found in source texts and system translations [19]. Although these methods are simple and interpretable, they suffer from the coverage of lexical alignments.

The second direction is embedding-based methods that align word or sentence embeddings between source sentences and system translations. YiSi-2 evaluates system translations by summing similarity scores over words pairs which are best-aligned mutual translations [16]. Moreover, by introducing cross-lingual linear projection, Lo and Larkin greatly improved the effect of YiSi-2 [17]. However, the cross-lingual linear projection is trained on limited sub-word token pairs and polysemy tokens will affect the projection effect. Zhao et al. proposed MoverScore metric using n-gram

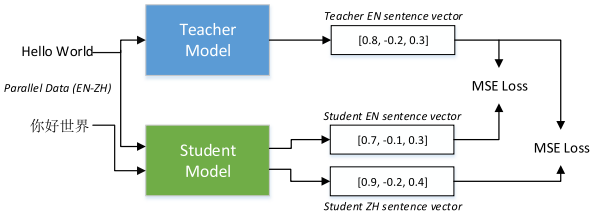

FIGURE 1. Procedure of Multilingual knowledge distillation (MKD) for sentence embedding alignment [21].

contextualized embeddings and Earth Mover Distance [18]. To mitigate the misalignment of cross-lingual word embedding spaces, Zhao et al. proposed XMoverScore metric which is a cross-lingual extension of MoverScore with post-hoc realignment strategies [28]. Song et al. proposed an unsupervised metric SentSim by incorporating a notion of sentence semantic similarity [20]. How to get high quality aligned word embeddings is crucial for these methods. Although directly aligning cross-lingual word embeddings is proposed in [31], the target language embeddings are constrained by their initialization, and the word pairs required for training are often noisy when they are obtained through unsupervised methods.

The third direction is model-based methods that train score models with annotated data. COMET-QE encodes segment-level representations of source and translation texts as the input to a feed forward regressor [27], [29]. UniTE is an unified framework engaged with abilities to handle reference-based and reference-free MT evaluations [30]. These methods work well if high-quality annotated data is sufficient. However, high-quality annotated data is still scarce in the field of reference-free MT evaluation [32].

In this paper, we follow the second direction and find out that the word embedding alignment could be implicitly achieved by sentence embedding alignment, where MKD can directly make use of large-scale parallel sentences and the contextual word embedding can alleviate the polysemy problem.

III. METHODOLOGY

A. MKD

The procedure of MKD proposed in [21] for sentence embedding alignment is shown in Fig. 1, where the teacher model is a monolingual SBERT [33] model, the student model is a multilingual pretrained model like mBERT or XLM-R, and the training data is composed of parallel sentences. The training procedure is to map the sentence embeddings of source and target sentences in parallel data that are obtained through the student model to the same location in the vector space as the source sentence embedding that is obtained through the teacher model by means of the MSE loss.

In this paper, we highlight that the word embedding alignment can be achieved implicitly by MKD. A theoretical analysis is first provided under a simplified condition.

Supposing $\boldsymbol{s}$ and $\boldsymbol{s}'$ are two source sentences, and $\boldsymbol{t}$ and $\boldsymbol{t}'$ are the corresponding target sentences. In the simplified

TABLE 1. Results of similarity metrics *SS* and *ST* for different layers (3rd, 6th, 9th and 12th) of model pmmb-v2 on 8 language pairs of WMT19 (Larger is better).

Metric	Layer	de-en	fi-en	gu-en	zh-en	en-cs	en-ru	en-zh	fr-de
SS	3	0.530	0.426	0.309	0.460	0.491	0.478	0.430	0.505
	6	0.708	0.669	0.521	0.660	0.703	0.697	0.635	0.726
	9	0.795	0.775	0.624	0.760	0.800	0.788	0.743	0.824
	12	**0.858**	**0.847**	**0.707**	**0.829**	**0.864**	**0.839**	**0.811**	**0.895**
ST	3	0.559	0.446	0.352	0.422	0.459	0.446	0.429	0.520
	6	0.723	0.671	0.550	0.627	0.690	0.676	0.629	0.736
	9	0.801	0.771	0.648	0.732	0.798	0.776	0.736	0.829
	12	**0.853**	**0.836**	**0.708**	**0.804**	**0.870**	**0.837**	**0.812**	**0.897**

condition, $\boldsymbol{s}$ and $\boldsymbol{t}$ are supposed to be the prefix substrings of $\boldsymbol{s'}$ and $\boldsymbol{t'}$ respectively. Then after tokenization, the four sentences could be represented as:

$$\boldsymbol{s} = (s_1, \ldots, s_m), \boldsymbol{s'} = (s_1, \ldots, s_m, \ldots, s_n), \tag{1}$$

$$\boldsymbol{t} = (t_1, \ldots, t_m), \boldsymbol{t'} = (t_1, \ldots, t_m, \ldots, t_n). \tag{2}$$

According to the mean pooling strategy used in SBERT [33] and MKD [21], the sentence embedding is the average of all token embeddings in the last layer of the given model. For sentence $\boldsymbol{s}$, the sentence embedding (SE) is:

$$SE(\boldsymbol{s}) = \frac{1}{m}\sum_{i=1}^{m} E_{LL}(s_i), \tag{3}$$

where $E_{LL}(s_i)$ stands for the contextual word embedding of s_i in the last layer (LL).

According to the training strategy in MKD, the sentence embeddings in the student model (after distillation) satisfy:

$$SE(\boldsymbol{s}) \approx SE(\boldsymbol{t}), SE(\boldsymbol{s'}) \approx SE(\boldsymbol{t'}), \tag{4}$$

i.e.,

$$\frac{1}{m}\sum_{i=1}^{m} E_{LL}(s_i) \approx \frac{1}{p}\sum_{i=1}^{p} E_{LL}(t_i), \tag{5}$$

$$\frac{1}{n}\sum_{i=1}^{n} E_{LL}(s_i) \approx \frac{1}{q}\sum_{i=1}^{q} E_{LL}(t_i). \tag{6}$$

If $m \approx p$ and $n \approx q$, we could have: $\sum_{i=1}^{m} E_{LL}(s_i) \approx \sum_{i=1}^{p} E_{LL}(t_i)$, $\sum_{i=1}^{m} E_{LL}(s_i) \approx \sum_{i=1}^{p} E_{LL}(t_i)$, then we can deduce:

$$\sum_{i=m+1}^{n} E_{LL}(s_i) \approx \sum_{j=p+1}^{q} E_{LL}(t_j). \tag{7}$$

And when $n - m = 1$ and $q - p = 1$, it could be drawn:

$$E_{LL}(s_n) \approx E_{LL}(t_q), \tag{8}$$

which means the cross-lingual word embedding alignment could be achieved implicitly through MKD.

Although the analysis is simple, we could still infer that the last layer has the best cross-lingual word embedding alignment among all the layers in the student model. This is different from mBERT and XLM-R models, where the best layer is the 9th among all the 12 layers for reference-based MT evaluation with BERTScore [2].

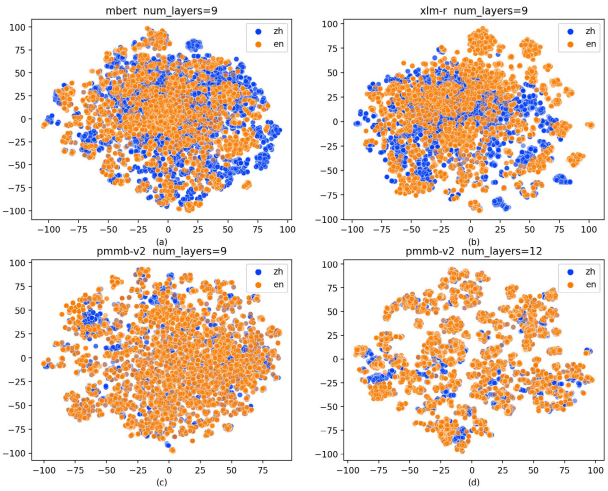

FIGURE 2. First two principle components of contextual token embeddings of mBERT (9th layer), XLM-R (9th layer) and pmmb-v2 (9th and 12th layers) for 100 zh-en parallel sentences in WMT19 by t-SNE (The more areas that do not cover each other, the worse the word embedding alignment effectiveness).

In order to illustrate the alignment effect intuitively, we design an example to compare the distilled student model with classic multilingual pretrained models mBERT [5] and XLM-R [10]. We choose the pretrained model paraphrase-multilingual-mpnet-base-v2[1] (distilled from XLM-R) as the student model (hereinafter referred to as pmmb-v2), and show the comparison results in Fig. 2.

In Fig. 2, each point represents a word in 100 zh-en parallel sentences from the WMT19 news translation shared task [13] and is composed of the first two principle components of the word embeddings of the respective models by t-SNE [34]. Because each word could be well aligned in the high-quality parallel sentences, the points representing the two language words will be covered by each other if no misalignment exists in the cross-lingual embedding spaces. From Fig. 2, it could

[1] more details in https://www.sbert.net/docs/pretrained_models.html

TABLE 2. *SS* values and source sentence ratios on different word count intervals for 8 language pair (LP) datasets from WMT19. (0, 10] represents the source sentences with the numbers of words in this interval, and so on for others.

LP	$(0, 10]$		$(10, 20]$		$(20, 30]$		$(30, 40]$		$(40, 50]$		$(50, +\infty)$	
	SS	Ratio	SS	Ratio	SS	Ratio	SS	Ratio	SS	Ratio	SS	Ratio
de-en	0.863	9.35%	0.850	31.70%	0.857	27.90%	0.866	17.15%	0.859	8.75%	0.857	5.15%
fi-en	0.835	8.92%	0.847	37.07%	0.846	30.86%	0.850	15.83%	0.849	5.46%	0.841	1.85%
gu-en	0.716	4.82%	0.716	31.20%	0.703	32.78%	0.697	15.35%	0.712	8.66%	0.714	7.19%
zh-en	0.848	2.40%	0.837	21.95%	0.837	24.30%	0.838	18.95%	0.829	12.75%	0.819	19.65%
en-cs	0.832	6.71%	0.864	24.39%	0.864	24.89%	0.864	21.68%	0.866	13.97%	0.866	8.36%
en-ru	0.850	6.71%	0.846	24.39%	0.842	24.89%	0.834	21.68%	0.833	13.97%	0.843	8.36%
en-zh	0.800	6.71%	0.808	24.39%	0.815	24.89%	0.812	21.68%	0.812	13.97%	0.807	8.36%
fr-de	0.875	4.17%	0.888	20.40%	0.898	23.87%	0.898	21.75%	0.896	12.99%	0.892	16.81%

be clearly discovered that the misalignment areas in the parts (c) and (d) for pmmb-v2 are much smaller than the parts (a) and (b) for mBERT and XLM-R. That is to say, MKD benefits cross-lingual word embedding alignment.

However, as shown in Fig. 2, part (d) (last layer) has better cluster properties than part (c) (9th layer), but it is not obvious that part (d) has a better alignment effect than part (c). In order to compare the word alignment effects of different layers in the student model pmmb-v2 quantitatively, we compute two similarity metrics on the word level for the set of high-quality parallel sentences $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$, which are defined:

$$SS = \sum_{\langle \boldsymbol{x},\boldsymbol{y} \rangle} \sum_{x_i \in \boldsymbol{x}} \max_{y_j \in \boldsymbol{y}} E_L(x_i)^\top E_L(y_j) \;/\; \sum_{\langle \boldsymbol{x},\boldsymbol{y} \rangle} |\boldsymbol{x}|, \tag{9}$$

$$ST = \sum_{\langle \boldsymbol{x},\boldsymbol{y} \rangle} \sum_{y_j \in \boldsymbol{y}} \max_{x_i \in \boldsymbol{x}} E_L(y_j)^\top E_L(x_i) \;/\; \sum_{\langle \boldsymbol{x},\boldsymbol{y} \rangle} |\boldsymbol{y}|, \tag{10}$$

where E_L is the cross-lingual word embedding in the selected layer (L) for a given token, and $|\boldsymbol{x}|$ and $|\boldsymbol{y}|$ denote the token numbers in sentences $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively. From the above definition, it could be seen that the two metrics *SS* and *ST* measure the degree of word alignment in source and target sentences (the larger the better)

We calculate the *SS* and *ST* for the 3rd, 6th, 9th and 12th layers of model pmmb-v2 (a total of 12 layers) on 8 language pair datasets from WMT19 (de-en, fi-en, gu-en, zh-en, en-cs, en-ru, en-zh and fr-de), and the results are illustrated in Table 1. From Table 1, it is obvious that the last layer of pmmb-v2 has the best results on all selected language pairs, which is consistent with our analysis. And we also find that the layer closer to the last layer has better results, which is very intuitive.

Since the word embedding alignment is achieved by sentence embedding alignment, the number of words in a sentence would affect the alignment effectiveness. Therefore, we report the *SS* values with different word count intervals for 8 language pair datasets from WMT19 in Table 2 (the 12th layer of model pmmb-v2 is used for de-en, fi-en, gu-en, zh-en, en-cs, en-ru, en-zh and fr-de). We divide the parallel sentences into 6 intervals according to the number of words in source sentences ((0, 10], (10, 20], (20, 30], (30, 40], (40, 50] and (50, $+\infty$)), and calculate the *SS* value for each interval. From Table 2, although the word embedding alignment effect somewhat degrades with the increase of words in sentences, it is still very robust because the *SS* values do not change much.

In short, through the above analysis and case studies on pmmb-v2, it could be seen that MKD could achieve cross-lingual word embedding alignment implicitly. Although the effectiveness of MKD is validated by the experimental results on reference-free MT evaluation in this paper, the reason why MKD could achieve cross-lingual word embedding alignment is still very worthy of in-depth analysis.

B. BERTScore

BERTScore [2] is an effective and robust automatic evaluation metric for text generation, which uses contextual embeddings to compute a similarity score for each token in the candidate sentence $\hat{\boldsymbol{x}}$ with each token in the reference sentence $\boldsymbol{x}$. In the absence of token importance weighting, the recall R, precision P and $F1$ score are defined as:

$$R = \frac{1}{|\boldsymbol{x}|} \sum_{x_i \in \boldsymbol{x}} \max_{\hat{x}_j \in \hat{\boldsymbol{x}}} E(x_i \mid \boldsymbol{x})^\top E(\hat{x}_j \mid \hat{\boldsymbol{x}}), \tag{11}$$

$$P = \frac{1}{|\hat{\boldsymbol{x}}|} \sum_{\hat{x}_j \in \hat{\boldsymbol{x}}} \max_{x_i \in \boldsymbol{x}} E(\hat{x}_j \mid \hat{\boldsymbol{x}})^\top E(x_i \mid \boldsymbol{x}), \tag{12}$$

$$F1 = 2\frac{P \cdot R}{P + R}, \tag{13}$$

where E is the contextual word embedding for a given token, the outputs of E are normalized to reduce similarity computation, and x_i and $\hat{x}_j$ denote the i-th and j-th tokens in $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$ respectively.

For MT evaluation, BERTScore with a pretrained model is usually used as a reference-based metric, which demonstrates stronger correlations with human judgments than BLEU. It is shown in this paper that BERTScore with the distilled student model is also suitable as a reference-free metric.

C. WMD

Word Mover's Distance (WMD) [22] measures the dissimilarity between two text documents as the minimum amount of distance that the embedded words of one document need to "travel" to reach the embedded words of the other document. WMD has been proven to generate high-quality results for

various text similarity tasks. Zhao et al. combined contextual embeddings and WMD as MoverScore [18] and proposed the extension XMoverScore [28] for text generation evaluation. Although it is observed that using WMD does not consistently improve BERTScore for reference-base MT evaluation [2], WMD is also chosen as a reference-free metric when the cross-lingual word embeddings of the distilled student model are used.

D. TARGET-SIDE LANGUAGE MODEL

As we know, faithfulness and fluency are two fundamental factors of translation quality [35]. In this paper, the alignment of cross-lingual word embeddings is mainly for the faithfulness of system translations; the target-side language model is for fluency, similar to the way used in [28]. A little different from the original sentence perplexity calculation [36], the calculation proposed in [23] for the pretrained language models like mBERT or XLM-R is used:

$$PPL(\boldsymbol{y}) = \frac{1}{|\boldsymbol{y}|}\sum_{y_i}\log\frac{1}{P(y_i|\boldsymbol{y}-y_i)}, \tag{14}$$

where PPL is the perplexity for given sentence $\boldsymbol{y}$, and $P(y_i|\boldsymbol{y}-y_i)$ is the probability of token y_i predicted by the pretrained language model when y_i is replaced by `[MASK]` in sentence $\boldsymbol{y}$.

E. METRICS

The cross-lingual word embeddings of the distilled student model are used in the framework of BERTScore or WMD as metrics for reference-free MT evaluation. In order to further improve the performance, a linear combination of BERTScore (or WMD) and the target-side language model (LM) is introduced.

For a source sentence $\boldsymbol{x}$ and a system translation sentence $\boldsymbol{y}$, the combined score (cs) for BERTScore is:

$$cs = (1-\lambda)\cdot BERTScore_{F1}(\boldsymbol{x},\boldsymbol{y}) - \lambda\cdot PPL(\boldsymbol{y}), \tag{15}$$

and the score for WMD is:

$$cs = (1-\lambda)\cdot WMD(\boldsymbol{x},\boldsymbol{y}) + \lambda\cdot PPL(\boldsymbol{y}), \tag{16}$$

where λ is a hyper-parameter and the values of BERTScore, WMD and PPL are normalized before combination (using a min-max normalization function $f(z) = (z-a)/(b-a)$).

It should be pointed out that minus is used in the combination for BERTScore because the smaller PPL is better (same for WMD), and the opposite for BERTScore.

IV. EXPERIMENT SETUP

In this section, we evaluate the performances of our above metrics by correlating them with human judgments of translation quality for reference-free MT evaluations, where both segment-level and system-level evaluations are included for full comparisons and are defined as follows.

Segment-level metrics (the input is a source sentence and a system translation sentence): The outputs of the last layer in the model pmmb-v2 are chosen as the cross-lingual word embeddings. We denote BERTScore[2] and WMD[3] that use those word embeddings as MKD-BERTScore and MKD-WMD metrics respectively. And the combinations with the target-side language model are denoted as MKD-BERTScore+LM and MKD-WMD+LM metrics, where the parameters (a, b) in normalization function $f(z)$ are set to (0.5, 1.0), (0.0, 2.0) and (1.0, 3.0) for MKD-BERTScore, MKD-WMD and LM respectively.

System-level metrics (the input is a set of source sentences and the corresponding system translation sentences): The mean values of our segment-level metrics on each pair of the sentences are used as the scores of our system-level metrics.

A. DATASETS

The source language sentences, and their system and reference translations are collected from the WMT19 news translation shared tasks [13], which contain predictions of 233 translation systems across 18 language pairs.[4] Each language pair in WMT19 has about 3,000 source sentences, and each source sentence is associated with one reference translation and with the automatic translations generated by participating systems. All 18 language pairs in WMT19, including 3 types of language directions (into-English (xx2en), from-English (en2xx), and none-English (xx2xx)), are evaluated in this paper.

B. BASELINES

From the above descriptions of our four metrics (MKD-BERTScore, MKD-WMD, MKD-BERTScore+LM and MKD-WMD+LM), it could be seen that these metrics do not require specific annotated data for MT evaluation, which means that these metrics are unsupervised embedding-based methods. Therefore, a range of reference-free metrics that are unsupervised or embedding-based are chosen for fair comparison: UNI and UNI+ [13], YiSi-2 [16] and YiSi-2+CLP [17], KoBE [19], XMoverScore [28], Prism-src [26]. To the best of our knowledge, the above metrics could cover most of the current SOTA metrics for reference-free MT evaluation. In addition, BERTScore that uses XLM-R[5] is denoted as BERTScore+XLM-R (the word embeddings in the 9th layer are used accroding to [2]) and is selected to directly compare the cross-lingual word embedding alignment effect with our metric MKD-BERTScore; and reference-based baseline metrics BLEU and sentBLEU [37] are selected as references. It should be pointed out that only the results of our metrics and BERTScore+XLM-R are calculated in this paper, and the results of the other metrics are from their respective papers.

C. EVALUATION MEASURES

Pearson correlation (r) and Kendall's Tau correlation (τ) [13] are used as measures for system-level and segment-level

[2]https://github.com/Tiiiger/bert_score
[3]https://github.com/src-d/wmd-relax
[4]https://github.com/AIPHES/ACL20-Reference-Free-MT-Evaluation
[5]https://huggingface.co/xlm-roberta-base

TABLE 3. Segment-level metric results for into-English language pairs of WMT19: absolute Kendall's Tau correlation of segment-level metric scores with DA. Best results excluding sentBLEU are in bold.

Metrics	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en	Avg
sentBLEU	0.056	0.233	0.188	0.377	0.262	0.125	0.323	0.223
UNI	0.022	0.202	-	-	-	0.084	-	-
UNI+	0.015	0.211	-	-	-	0.089	-	-
YiSi-2	0.068	0.126	-0.001	0.096	0.075	0.053	0.253	0.096
YiSi-2+CLP	0.116	0.271	**0.249**	0.370	0.281	0.121	**0.340**	0.250
Prism-src	0.109	0.300	0.102	**0.391**	**0.356**	**0.178**	0.336	0.253
BERTScore+XLM-R	0.084	0.185	0.149	0.176	0.144	0.057	0.157	0.136
MKD-BERTScore	0.093	0.234	0.171	0.310	0.211	0.089	0.208	0.188
MKD-BERTScore+LM	0.129	0.294	0.226	0.333	0.290	0.124	0.282	0.240
MKD-WMD	0.094	0.233	0.175	0.310	0.206	0.080	0.214	0.188
MKD-WMD+LM	**0.139**	**0.307**	0.248	0.342	0.307	0.130	0.302	**0.254**

TABLE 4. Segment-level metric results for from-English language pairs of WMT19: absolute Kendall's Tau correlation of segment-level metric scores with DA. Best results excluding sentBLEU are in bold.

Metrics	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh	Avg
sentBLEU	0.367	0.248	0.396	0.465	0.392	0.334	0.469	0.270	0.368
UNI	0.060	0.129	0.351	-	-	-	0.226	-	-
UNI+	-	-	-	-	-	-	0.222	-	-
YiSi-2	0.069	0.212	0.239	0.147	0.187	0.003	-0.155	0.044	0.093
YiSi-2+CLP	0.299	0.329	0.459	**0.512**	0.459	0.314	0.078	0.158	0.326
Prism-src	**0.470**	**0.402**	**0.555**	0.215	**0.507**	**0.499**	**0.486**	0.287	**0.428**
BERTScore+XLM-R	0.045	0.204	0.224	0.289	0.253	0.013	-0.151	0.040	0.115
MKD-BERTScore	0.151	0.284	0.357	0.326	0.280	0.179	-0.065	0.085	0.200
MKD-BERTScore+LM	0.461	0.395	0.531	0.438	0.433	0.489	0.307	**0.324**	0.422
MKD-WMD	0.133	0.278	0.342	0.335	0.294	0.165	-0.079	0.087	0.194
MKD-WMD+LM	0.455	0.374	0.518	0.389	0.363	0.493	0.306	0.309	0.401

metric evaluations respectively. Pearson correlation is defined as:

$$r = \frac{\sum_{i=1}^{n}(H_i - \overline{H})(M_i - \overline{M})}{\sqrt{\sum_{i=1}^{n}(H_i - \overline{H})^2} \cdot \sqrt{\sum_{i=1}^{n}(M_i - \overline{M})^2}}, \tag{17}$$

where H_i are human assessment scores of all systems (or sentence pairs) in a given translation direction, M_i are the corresponding scores predicted by a given metric, and $\overline{H}$ and $\overline{M}$ are their mean values respectively.

And Kendall's Tau correlation is defined as:

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|}, \tag{18}$$

where *Concordant* is the set of all human comparisons for which a given metric suggests the same order, and *Discordant* is the set of all human comparisons with which a given metric disagrees.

V. RESULTS

A. MAIN RESULTS

Tables 3-5 and Tables 6-8 show the comparison results of segment-level and system-level evaluations on the 18 language pairs of WMT19 respectively, where the hyper-parameter λ for our metrics MKD-BERTScore+LM and MKD-WMD+LM is set to 0.1 on the xx2en and xx2xx language directions, and set to 0.4 on the en2xx language direction.

TABLE 5. Segment-level metric results for none-English language pairs of WMT19: absolute Kendall's Tau correlation of segment-level metric scores with DA. Best results excluding sentBLEU are in bold.

Metrics	de-cs	de-fr	fr-de	Avg
sentBLEU	0.203	0.235	0.179	0.206
YiSi-2	0.199	0.186	0.066	0.150
YiSi-2+CLP	0.355	0.294	0.226	0.292
Prism-src	**0.444**	**0.374**	**0.312**	**0.377**
BERTScore+XLM-R	0.201	0.184	0.049	0.145
MKD-BERTScore	0.412	0.231	0.186	0.276
MKD-BERTScore+LM	0.442	0.295	0.238	0.325
MKD-WMD	0.407	0.241	0.170	0.272
MKD-WMD+LM	0.439	0.308	0.232	0.326

From the comparison results of BERTScore+XLM-R and MKD-BERTScore metrics, it could be seen that MKD-BERTScore has significantly better results on all language pairs for both segment-level (avg. xx2en 0.136 → 0.188, en2xx 0.115 → 0.200, xx2xx 0.145 → 0.276) and system-level (avg. xx2en 0.396 → 0.806, en2xx 0.238 → 0.469, xx2xx 0.173 → 0.763) evaluations, which indicates

TABLE 6. System-level metric results for into-English language pairs of WMT19: absolute Pearson correlation of system-level metric scores with DA. Best results excluding BLEU are in bold.

Metrics	de-en	fi-en	gu-en	kk-en	lt-en	ru-en	zh-en	Avg
BLEU	0.849	0.982	0.834	0.946	0.961	0.879	0.899	0.907
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	0.578
YiSi-2+CLP	**0.898**	**0.959**	0.739	0.981	0.935	0.461	0.980	0.850
KoBE	0.863	0.538	**0.828**	0.899	0.704	**0.928**	0.907	0.810
XMoverScore	0.625	0.890	-0.060	**0.993**	0.851	**0.928**	0.968	0.742
Prism-src	0.890	0.941	0.171	0.961	**0.989**	0.845	0.971	0.824
BERTScore+XLM-R	0.785	0.866	-0.007	0.117	0.657	-0.372	0.728	0.396
MKD-BERTScore	0.823	0.956	0.420	0.828	0.946	0.747	0.924	0.806
MKD-BERTScore+LM	0.888	0.951	0.579	0.941	0.978	0.851	**0.991**	0.883
MKD-WMD	0.818	**0.959**	0.465	0.849	0.941	0.288	0.914	0.748
MKD-WMD+LM	0.894	0.947	0.684	0.984	0.977	0.761	0.943	**0.884**

TABLE 7. System-level metric results for from-English language pairs of WMT19: absolute Pearson correlation of system-level metric scores with DA. Best results excluding BLEU are in bold.

Metrics	en-cs	en-de	en-fi	en-gu	en-kk	en-lt	en-ru	en-zh	Avg
BLEU	0.897	0.921	0.969	0.737	0.852	0.989	0.986	0.901	0.907
UNI	0.028	0.841	0.907	-	-	-	**0.919**	-	-
UNI+	-	-	-	-	-	-	0.918	-	-
YiSi-2	0.324	0.924	0.696	0.314	0.339	0.055	0.766	0.097	0.439
YiSi-2+CLP	0.773	0.963	0.906	**0.890**	0.977	0.761	0.473	0.449	0.774
KoBE	0.597	0.888	0.521	-0.340	0.827	-0.049	0.895	0.216	0.444
Prism-src	0.865	0.976	0.933	0.444	0.959	0.908	0.822	0.793	0.838
BERTScore+XLM-R	0.035	0.893	0.765	0.549	0.650	-0.084	-0.779	-0.127	0.238
MKD-BERTScore	0.585	0.948	0.844	0.612	0.567	0.501	-0.473	0.164	0.469
MKD-BERTScore+LM	0.895	**0.985**	**0.945**	0.809	0.976	0.904	0.662	0.944	**0.890**
MKD-WMD	0.498	0.944	0.833	0.621	0.636	0.442	-0.704	0.119	0.424
MKD-WMD+LM	**0.896**	0.984	**0.945**	0.740	0.931	**0.917**	0.653	**0.955**	0.878

the cross-lingual word embeddings by MKD have much better alignment effect because only the word embeddings are different for the two metrics.

And when being compared with the current SOTA metrics involved in this paper for both segment-level and system-level evaluations in WMT19, with the assistant of the target-side language model, our metrics get 4 best average results on all 3 types of language directions (avg. segment-level 0.254 for xx2en, and system-level (0.884, 0.890, 0.898) for xx2en, en2xx, xx2xx) and rank first on more than half of the 18 language pairs for system-level evaluations (11/18).

Prism-src is a very competitive unsupervised metric, which frames the task of MT evaluation as one of scoring machine translation output with a sequence-to-sequence paraphraser conditioned on source text [26]. The results in Table 7 and Table 8 show that our metrics have better performances than Prism-src on the system-level evaluations of the from-English and none-English language pairs, although they do not outperform Prism-src on the segment-level evaluations as illustrated in Table 4 and Table 5.

TABLE 8. System-level metric results for none-English language pairs of WMT19: absolute Pearson correlation of system-level metric scores with DA. Best results excluding BLEU are in bold.

Metrics	de-cs	de-fr	fr-de	Avg
BLEU	0.941	0.891	0.864	0.899
YiSi-2	0.606	0.721	0.530	0.619
YiSi-2+CLP	0.860	0.853	0.461	0.725
Prism-src	0.973	0.889	0.739	0.867
BERTScore+XLM-R	0.572	0.692	-0.746	0.173
MKD-BERTScore	0.979	0.826	0.483	0.763
MKD-BERTScore+LM	0.979	0.892	**0.809**	0.893
MKD-WMD	**0.983**	0.823	0.259	0.688
MKD-WMD+LM	0.978	**0.914**	0.802	**0.898**

Therefore, our metrics are very competitive for reference-free MT evaluation. In addition, it is worth mentioning that BERTScore and WMD are almost the same in our metrics for reference-free MT evaluation, which is consistent with the conclusion in [2] for reference-based MT evaluation.

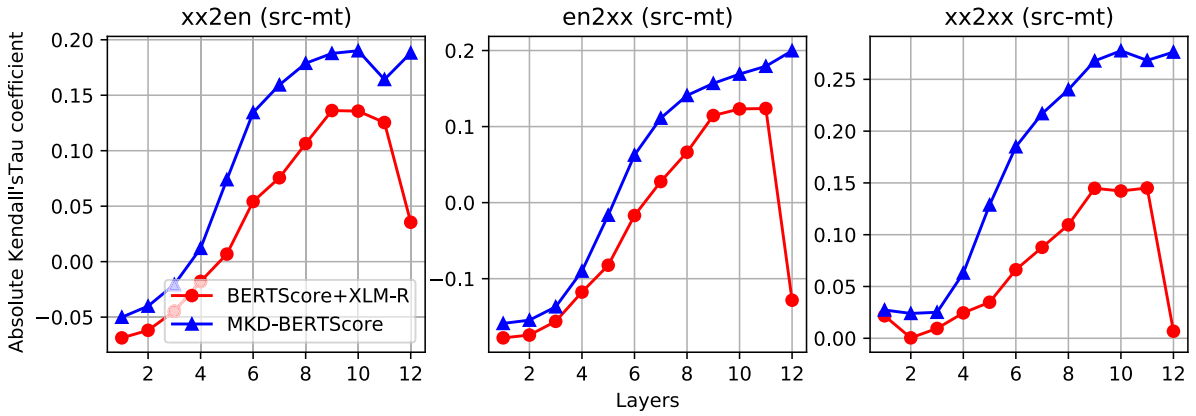

FIGURE 3. Mean absolute Kendall's Tau correlation of MKD-BERTScore and BERTScore+XLM-R with different layers of word embeddings for segment-level reference-free MT evaluation on WMT19.

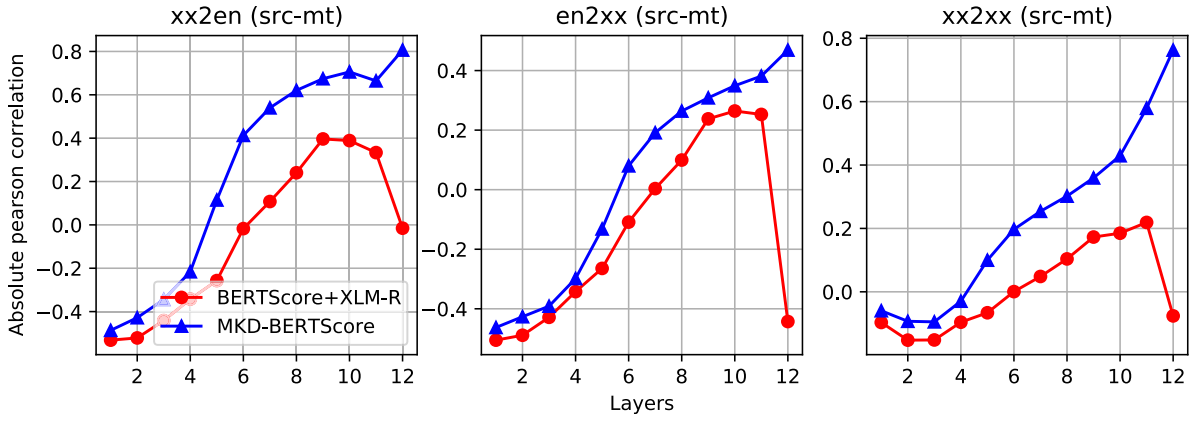

FIGURE 4. Mean absolute Pearson correlation of MKD-BERTScore and BERTScore+XLM-R with different layers of word embeddings for system-level reference-free MT evaluation on WMT19.

B. EFFECT OF EMBEDDING LAYERS

BERTScore is sensitive to the layer of the model selected to generate the contextual token embeddings [2]. We still investigate which layer is the best choice for the model pmmb-v2 on WMT19 through experimental comparisons, although the last layer is theoretically proved to be the best under the simplified condition. The metric MKD-BERTScore is selected for investigation, and BERTScore+XLM-R is chosen for comparison. The mean values on the into-English, from-English and none-English language pairs of WMT19 for segment-level and system-level evaluations are illustrated in Fig. 3 and Fig. 4 respectively.

From Fig. 3 and Fig. 4, it could be clearly seen that the last layer is indeed the best choice for MKD-BERTScore on both segment-level and system-level evaluations, which is fully consistent with our theoretical analysis. And it is interesting to find that the best layers of BERTScore+XLM-R for reference-free and reference-based evaluations are almost the same (9th). Meanwhile, MKD-BERTScore outperforms BERTScore+XLM-R on every layer for both segment-level and system-level evaluations.

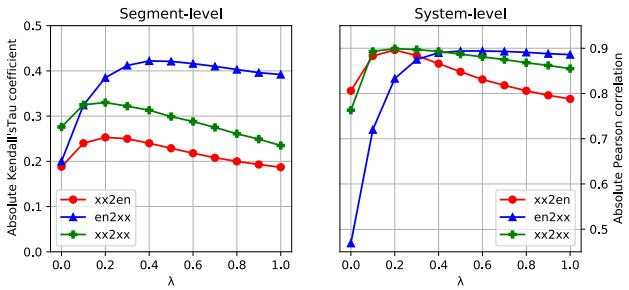

FIGURE 5. Mean absolute Kendall's Tau correlation for segment-level and mean absolute Pearson correlation for system-level with different combination parameter λ for MKD-BERTScore and LM on WMT19.

C. COMBINATION PARAMETER STUDY

In this section, the combination parameter λ for MKD-BERTScore and LM is studied, which varies from 0 to 1 with step 0.1. Fig. 5 shows the mean evaluation values on the into-English, from-English and none-English language pairs of WMT19 for segment-level and system-level evaluations with different values of the parameter λ.

TABLE 9. Kendall's Tau correlation of the metrics (MKD-BERTScore, LM, MKD-BERTScore+LM) in the experiments on the zh-en dataset from WMT19 (Experiment 1: *faithful-but-not-fluent*, Experiment 2: *fluent-but-not-faithful*).

Metrics	Experiment 1	Experiment 2
MKD-BERTScore	0.531	**1.000**
LM	**0.974**	0.006
MKD-BERTScore+LM	0.953	**1.000**

From Fig. 5, it is obvious that the combination indeed improves the performance of our metrics on both segment-level and system-level evaluations. For the into-English and none-English language pairs, a smaller value for λ (such as 0.1) is a good choice, while a relative larger value (about 0.4) is suitable for the from-English language pairs. This is because the model pmmb-v2 is aligned with English for other languages, which deteriorates the pmmb-v2's language model capabilities for other languages. As a result, larger values are required for parameter λ to increase the importance of LM for other languages. In addition, it is worth noting that LM as a metric (i.e. $\lambda = 1$) could get very competitive results especially on the from-English language pairs, which means fluency of system translations is very important in MT evaluation.

D. DISCUSSION

In this section, we further analyze the impacts of faithfulness and fluency to the metrics (MKD-BERTScore, LM, MKD-BERTScore+LM) by designing the following experiments: (1) *faithful-but-not-fluent*: randomly shuffle 2 to 5 continuous words for each translation sentence; (2) *fluent-but-not-faithful*: randomly shuffle translation sentences, i.e., the translation of another source sentence is used as the translation of the current source sentence. The zh-en dataset from WMT19 is chosen for experiments, which contains 2,000 parallel sentences. For each source sentence, we can get a correct translation and a shuffled translation in each experiment, which can be composed into a triplet <*source sentence, correct translation, shuffled translation*>. So the Kendall's Tau correlation τ can be used to evaluate the metrics on the designed experiments, and the results are illustrated in Table 9.

As shown in Table 9, the metric LM is very effective for the *faithful-but-not-fluent* experiment, but fails for the *fluent-but-not-faithful* experiment. Meanwhile, the metric MKD-BERTScore works very well for the *fluent-but-not-faithful* experiment, but not well for the *faithful-but-not-fluent* experiment. MKD-BERTScore+LM, as a fusion of the two metrics, can achieve very good results for the two experiments.

It should be noted that the metrics (MKD-BERTScore, LM, MKD-BERTScore+LM) may be not effective when the source sentences have noises, such as in simultaneous interpretation scenarios. This is because the pretrained models for these metrics are obtained on noise-free data and these metrics tend to give better scores to noise-free translations when the correct translations may have noises. To mitigate this problem, fine-tuning the pretrained models with noise data might be a solution.

VI. CONCLUSION

In this paper, it is revealed by a simplified theoretical analysis that the cross-lingual word embedding alignment could be achieved implicitly through MKD for sentence embedding alignment. And with the frameworks of BERTScore and WMD, the cross-lingual word embeddings are applied as metrics (MKD-BERTScore and MKD-WMD) for the reference-free MT evaluations. From the experimental results on WMT19, with linear combination of the target-side LM, our metrics could get 4 best average scores on all 3 types of language directions, and rank first on more than half of all the language pairs (11 out of 18) for system-level evaluations, when the current SOTA reference-free metrics that we know are selected for comparison. Meanwhile, it is proved by theoretical analysis and experimental comparisons that the last layer of the distilled model is the best choice for reference-free MT evaluation. The linear combination parameter λ for MKD-BERTScore and LM is also investigated. Nevertheless, the reason why MKD could achieve the alignment of cross-lingual word embeddings is still very worthy of further study.

REFERENCES

[1] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*. Philadelphia, PA, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040

[2] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTSCORE: Evaluating text generation with BERT," in *Proc. 8th Int. Conf. Learn. Represent. (ICLR)*, Addis Ababa, Ethiopia, Apr. 2020, pp. 1–43. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[3] T. Sellam, D. Das, and A. Parikh, "BLEURT: Learning robust metrics for text generation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, Jun. 2020, pp. 7881–7892.

[4] O. F. Zaidan and C. Callison-Burch, "Crowdsourcing translation: Professional quality from non-professionals," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics Human Lang. Technol.* Portland, OR, USA: Association for Computational Linguistics, Jun. 2011, pp. 1220–1229. [Online]. Available: https://aclanthology.org/P11-1122

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Language Technol.* Minneapolis, MN: Association for Computational Linguistics, vol. 1, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[6] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?" in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4996–5001. [Online]. Available: https://aclanthology.org/P19-1493

[7] M. Artetxe and H. Schwenk, "Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond," *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 597–610, Sep. 2019, doi: 10.1162/tacl_a_00288.

[8] A. Conneau and G. Lample, "Cross-lingual language model pretraining," in *Advances in Neural Information Processing Systems*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 7059–7069.

[9] M. Chidambaram, Y. Yang, D. Cer, S. Yuan, Y. Sung, B. Strope, and R. Kurzweil, "Learning cross-lingual sentence representations via a multi-task dual-encoder model," in *Proc. 4th Workshop Represent. Learn. (RepLNLP)*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 250–259. [Online]. Available:https://aclanthology.org/W19-4330

[10] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds. Stroudsburg, PA, USA: Association for Computational Linguistics, Jun. 2020, pp. 8440–8451, doi: 10.18653/v1/2020.acl-main.747.

[11] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic BERT sentence embedding," 2020, *arXiv:2007.01852*.

[12] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 726–742, Dec. 2020. [Online]. Available: https://transacl.org/ojs/index.php/tacl/article/view/2107

[13] Q. Ma, J. Wei, O. Bojar, and Y. Graham, "Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges," in *Proc. 4th Conf. Mach. Transl.* Florence, Italy: Association for Computational Linguistics, vol. 2, Aug. 2019, pp. 62–90. [Online]. Available: https://aclanthology.org/W19-5302

[14] M. Popović, D. Vilar, E. Avramidis, and A. Burchardt, "Evaluation without references: IBM1 scores as evaluation metrics," in *Proc. 6th Workshop Stat. Mach. Transl.* Edinburgh, Scotland: Association for Computational Linguistics, Jul. 2011, pp. 99–103. [Online]. Available: https://aclanthology.org/W11-2109

[15] L. Specia, K. Shah, J. G. de Souza, and T. Cohn, "QuEst—A translation quality estimation framework," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*. Sofia, Bulgaria: Association for Computational Linguistics, Aug. 2013, pp. 79–84. [Online]. Available: https://aclanthology.org/P13-4014

[16] C.-K. Lo, "YiSi—A unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources," in *Proc. 4th Conf. Mach. Transl.* Florence, Italy: Association for Computational Linguistics, vol. 2, Aug. 2019, pp. 507–513. [Online]. Available: https://aclanthology.org/W19-5358

[17] C.-K. Lo and S. Larkin, "Machine translation reference-less evaluation using YiSi-2 with bilingual mappings of massive multilingual language model," in *Proc. 5th Conf. Mach. Transl.* Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2020, pp. 903–910. [Online]. Available: https://aclanthology.org/2020.wmt-1.100

[18] W. Zhao, M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger, "MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*. Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 563–578. [Online]. Available: https://aclanthology.org/D19-1053

[19] Z. Gekhman, R. Aharoni, G. Beryozkin, M. Freitag, and W. Macherey, "KoBE: Knowledge-based machine translation evaluation," in *Proc. Findings Assoc. Comput. Linguistics (EMNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2020, pp. 3200–3207. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.287

[20] Y. Song, J. Zhao, and L. Specia, "SentSim: Crosslingual semantic evaluation of machine translation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.* Stroudsburg, PA, USA: Association for Computational Linguistics, Jun. 2021, pp. 3143–3156. [Online]. Available: https://aclanthology.org/2021.naacl-main.252

[21] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2020, pp. 4512–4525. [Online]. Available: https://aclanthology.org/2020.emnlp-main.365

[22] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *Proc. 32nd Int. Conf. Mach. Learn.*, vol. 37, F. Bach and D. Blei, Eds. Lille, France: PMLR, Jul. 2015, pp. 957–966. [Online]. Available: https://proceedings.mlr.press/v37/kusnerb15.html

[23] M. Zhang, X. Qiao, H. Yang, S. Tao, Y. Zhao, Y. Li, C. Su, M. Wang, J. Guo, Y. Liu, and Y. Qin, "Target-side language model for reference-free machine translation evaluation," in *Machine Translation*, T. Xiao and J. Pino, Eds. Singapore: Springer Nature, 2022, pp. 45–53.

[24] M. Zhang, H. Yang, S. Tao, Y. Zhao, X. Qiao, Y. Li, C. Su, M. Wang, J. Guo, Y. Liu, and Y. Qin, "Incorporating multilingual knowledge distillation into machine translation evaluation," in *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers the Digital Economy*, M. Sun, G. Qi, K. Liu, J. Ren, B. Xu, Y. Feng, Y. Liu, and Y. Chen, Eds. Singapore: Springer Nature, 2022, pp. 148–160.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 6000–6010.

[26] B. Thompson and M. Post, "Automatic machine translation evaluation in many languages via zero-shot paraphrasing," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2020, pp. 90–121. [Online]. Available: https://aclanthology.org/2020.emnlp-main.8

[27] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "COMET: A neural framework for MT evaluation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*. Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2020, pp. 2685–2702. [Online]. Available: https://aclanthology.org/2020.emnlp-main.213

[28] W. Zhao, G. Glavaš, M. Peyrard, Y. Gao, R. West, and S. Eger, "On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, Jul. 2020, pp. 1656–1671. [Online]. Available: https://aclanthology.org/2020.acl-main.151

[29] R. Rei, A. C. Farinha, C. Zerva, D. van Stigt, C. Stewart, P. Ramos, T. Glushkova, A. F. T. Martins, and A. Lavie, "Are references really needed? Unbabel-IST 2021 submission for the metrics shared task," in *Proc. 6th Conf. Mach. Transl.*, Nov. 2021, pp. 1030–1040. [Online]. Available: https://aclanthology.org/2021.wmt-1.111

[30] Y. Wan, D. Liu, B. Yang, H. Zhang, B. Chen, D. Wong, and L. Chao, "UniTE: Unified translation evaluation," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8117–8127. [Online]. Available: https://aclanthology.org/2022.acl-long.558

[31] S. Cao, N. Kitaev, and D. Klein, "Multilingual alignment of contextual word representations," in *Proc. ICLR*, 2020, pp. 1–15. [Online]. Available: https://openreview.net/forum?id=r1xCMyBtPS

[32] M. Freitag, R. Rei, N. Mathur, C.-K. Lo, C. Stewart, E. Avramidis, T. Kocmi, G. Foster, A. Lavie, and A. F. T. Martins, "Results of WMT22 metrics shared task: Stop using BLEU—Neural metrics are better and more robust," in *Proc. 7th Conf. Mach. Transl.* Abu Dhabi: Association for Computational Linguistics, Dec. 2022, pp. 46–68. [Online]. Available: https://aclanthology.org/2022.wmt-1.2

[33] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.* Hong Kong: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992.

[34] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html

[35] P. Koehn, *Statistical Machine Translation*, 1st ed. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[36] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?" *Proc. IEEE*, vol. 88, no. 8, pp. 1270–1278, Aug. 2000. [Online]. Available: http://www.cs.cmu.edu/~roni/papers/survey-slm-IEEE-PROC-0004.pdf

[37] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, "Moses: Open source toolkit for statistical machine translation," in *Proc. 45th Annu. Meeting Assoc. Comput. Linguistics Companion*. Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 177–180. [Online]. Available: https://aclanthology.org/P07-2045

MIN ZHANG (Member, IEEE) received the B.S. and Ph.D. degrees in computer science and technology from the University of Science and Technology of China, in 2003 and 2008, respectively. He is currently an Expert with Huawei Translation Services Center in 2012 Lab, Huawei. His research interests include machine translation, knowledge graph, and natural language processing.

HAO YANG (Senior Member, IEEE) received the Ph.D. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2009. He is currently a Research Scientist with Huawei Translation Services Center in 2012 Lab, Huawei. His main research interests include machine translation and natural language processing.

YANQING ZHAO is currently an Expert with Huawei Translation Services Center in 2012 Lab, Huawei. Her main research interests include machine translation, knowledge graph, natural language processing.

XIAOSONG QIAO is currently a Research Engineer with Huawei Translation Services Center in 2012 Lab, Huawei. His main research interests include machine translation, knowledge graph, and natural language processing.

SHIMIN TAO received the M.S. degree from Beihang University, Beijing, China. He is currently an Expert with Huawei Translation Services Center in 2012 Lab, Huawei. His main research interests include machine translation and natural language processing.

SONG PENG is currently an Expert with Huawei Translation Services Center in 2012 Lab, Huawei. His main research interests include machine translation and natural language processing.

YING QIN is currently an Expert with Huawei Translation Services Center in 2012 Lab, Huawei. Her main research interests include machine translation and natural language processing.

YANFEI JIANG is currently an Expert with the 2012 Lab and the Director of the Huawei Translation Services Center, Huawei. His main research interests include machine translation and natural language processing.