Incorporating Multilingual Knowledge Distillation into Machine Translation Evaluation

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

Huawei Translation Services Center, Beijing, China
{zhangmin186,yanghao30,taoshimin,zhaoyanqing,qiaoxiaosong,liyinglu, suchang8,wangminghan,guojiaxin1,liuyilun3,qinying}@huawei.com

Abstract. Multilingual knowledge distillation is proposed for multilingual sentence embedding alignment. In this paper, it is found out that multilingual knowledge distillation could implicitly achieve cross-lingual word embedding alignment, which is critically important for reference-free machine translation evaluation (where source texts are directly compared with system translations). Then with the framework of BERTScore, we propose a metric BERTScore-MKD for reference-free machine translation evaluation. From the experimental results on the into-English language pairs of WMT17-19, the reference-free metric BERTScore-MKD is very competitive (not only best mean scores, but also better than BLEU on WMT17-18) when the current state-of-the-art (SOTA) metrics that we know are chosen for comparison. Moreover, the results on WMT19 demonstrate that BERTScore-MKD is also suitable for reference-based machine translation evaluation (where reference texts are used to be compared with system translations).

Keywords: Multilingual knowledge distillation · Machine translation evaluation · BERTScore-MKD

1 Introduction

In traditional machine translation (MT) evaluation (also referred to as *reference-based* MT evaluation), reference texts are provided and compared with system translations. The common metrics for such evaluation include the word-based metrics BLEU [1] and METEOR [2], and the word embedding-based metrics BERTScore [3] and BLEURT [4].

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 [5]. Recently, with the rapid progress of deep learning in multilingual language processing [6,7], there has been a growing interest in

M. Zhang and H. Yang—Equally contributed.

M. Sun et al. (Eds.): CCKS 2022, CCIS 1669, pp. 148–160, 2022.
https://doi.org/10.1007/978-981-19-7596-7_12

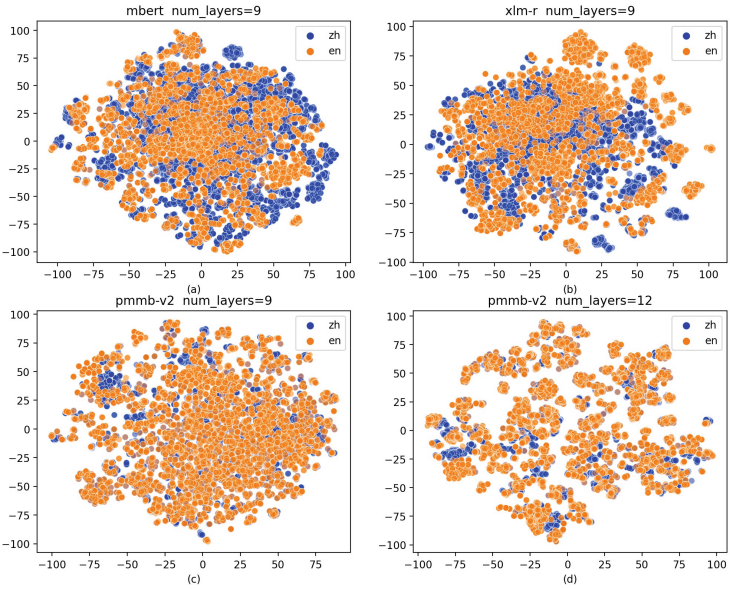

Fig. 1. First two principle components of contextual token embeddings of mBERT, XLM-R and pmmb-v2 for 100 zh-en parallel sentences in WMT19 by t-SNE (The more areas that do not cover each other, the worse the word embedding alignment effectiveness)

reference-free MT evaluation [8], which is also referred to as "quality estimation" (QE) in the MT community. In QE, evaluation metrics compare system translations with source sentences directly. And lots of methods have been proposed to approach this task. Popović et al. [9] exploited a bag-of-word translation model for quality estimation, which sums over the likelihoods of aligned word pairs between source and translation texts. Specia et al. [10] used language-agnostic linguistic features extracted from source texts and system translations to estimate quality. YiSi-2 [11] evaluates system translations by summing similarity scores over words pairs which are best-aligned mutual translations. Prism-src [12] frames the task of MT evaluation as one of scoring machine translation output with a sequence-to-sequence paraphraser, conditioned on source text. COMET-QE [13,14] encodes segment-level representations of source text and translation text as the input to a feed forward regressor. To mitigate the misalignment of cross-lingual word embedding spaces, Zhao et al. [15] proposed post-hoc realignment strategies which integrate a target-side GPT [16] language model. Song et al. [17] proposed an unsupervised metric SentSim by incorporating a notion of sentence semantic similarity. Wan et al. [18] proposed a unified framework (UniTE) with monotonic regional attention and unified pretraining for reference-only, source-only and source-reference-combined MT evaluations.

In a word, most of the above mentioned methods try to directly achieve cross-lingual alignment on lexical, word embedding or sentence embedding levels,

which is critically important for reference-free MT evaluation. In this paper, we find out that cross-lingual word embedding alignment could be achieved implicitly by multilingual knowledge distillation (MKD) [19] for sentence embedding alignment, of which the training procedure is to map the sentence embeddings of source and target sentences in parallel data that are obtained through a multilingual pretrained model to the same location in the vector space as the source sentence embedding that is obtained through a monolingual Sentence-BERT (SBERT) [20] model by means of the MSE loss. To illustrate the alignment effect intuitively, a simple example shown in Fig. 1 is designed to compare the distilled multilingual model (paraphrase-multilingual-mpnet-base-v2[1], hereinafter referred to as pmmb-v2) with the classic multilingual pretrained models mBERT [6] and XLM-R [7]. In Fig. 1, each point represents a word in 100 zh-en parallel sentences from the WMT19 news translation shared task [8] and is composed of the first two principle components of the contextual word embeddings of the respective models by t-SNE [21]. 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. 1, it could 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. This show that multilingual knowledge distillation benefits cross-lingual word embedding alignment.

In this paper, with the framework of BERTScore, we incorporate multilingual knowledge distillation into MT evaluation and propose a reference-free metric BERTScore-MKD. And then we test the performance of BERTScore-MKD on the into-English language pairs of WMT17-19 for both system-level and segment-level evaluations. The experimental results show that BERTScore-MKD is very competitive when compared with the current SOTA reference-free metrics that we know. Furthermore, from the comparison results on WMT19, it is interesting to find that BERTScore-MKD is also suitable for reference-based MT evaluation.

2 Method

In this section, the metric BERTScore-MKD will be given after the descriptions of multilingual knowledge distillation and BERTScore.

2.1 Multilingual Knowledge Distillation

The procedure of multilingual knowledge distillation (MKD) proposed by Reimers and Gurevych [19] for sentence embedding alignment is described in Fig. 2, where the teacher model is monolingual SBERT [20] which achieves state-of-the-art performance for various sentence embedding tasks, and the student model is a multilingual pretrained model like mBERT or XLM-R before distillation. From Fig. 2, it could be seen that MKD achieves the alignment of paired

[1] Distilled from XLM-R, more details in https://www.sbert.net/docs/pretrained_models.html.

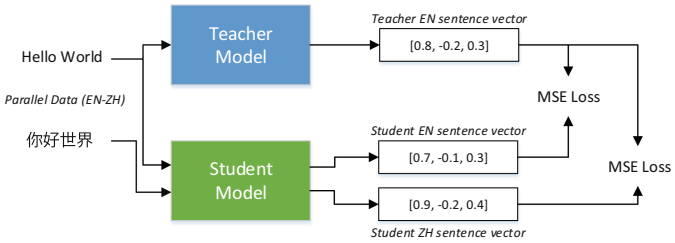

Fig. 2. Multilingual knowledge distillation [19]

sentence embedding directly. And the effectiveness of the student model's sentence embedding after distillation is demonstrated for over 50 languages from various language families [19].

2.2 BERTScore

BERTScore[2] [3] is an effective and robust automatic evaluation metric for text generation, which computes a similarity score for each token in the candidate sentence $\hat{\boldsymbol{x}}$ with each token in the reference sentence $\boldsymbol{x}$ by using contextual embedding instead of exact matches. In the absence of token importance weighting, the recall R, precision P and $F1$ score are defined as:

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

$$P = \frac{1}{|\hat{\boldsymbol{x}}|} \sum_{\hat{x}_j \in \hat{\boldsymbol{x}}} \max_{x_i \in \boldsymbol{x}} E(\hat{x}_j \mid \hat{\boldsymbol{x}})^\top E(x_i \mid \boldsymbol{x}), \tag{2}$$

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}, \tag{3}$$

where E is a contextual word embedding function, 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, and we will show that BERTScore using the distilled student model in Sect. 2.1 is suitable for both reference-free and reference-based MT evaluations.

2.3 BERTScore-MKD

Suppose $\boldsymbol{s}$ and $\boldsymbol{r}$ are two parallel sentences, which could be denoted as:

$$\boldsymbol{s} = (s_1, \ldots, s_i, \ldots, s_m), \tag{4}$$

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

$$\boldsymbol{r} = (r_1, \ldots, r_j, \ldots, r_n), \tag{5}$$

where s_i and r_j denote the i-th and j-th tokens in $\boldsymbol{s}$ and $\boldsymbol{r}$ respectively.

According to the mean pooling strategy used in SBERT and MKD [19,20], the sentence embedding is the average of all token embeddings in the last layer of the given model. So the two sentence embeddings of $\boldsymbol{s}$ and $\boldsymbol{r}$ for the student model could be represented as:

$$SE(\boldsymbol{s}) = \frac{1}{m}\sum_{i=1}^{m} E_{LL}(s_i \mid \boldsymbol{s}), \tag{6}$$

$$SE(\boldsymbol{r}) = \frac{1}{n}\sum_{j=1}^{n} E_{LL}(r_j \mid \boldsymbol{s}), \tag{7}$$

where SE denotes the sentence embedding of the given sentence, and E_{LL} stands for the contextual word embedding function in the last layer (LL).

As illustrated in Fig. 2, after distillation with MSE loss for the student model, we could have $SE(\boldsymbol{s}) \approx SE(\boldsymbol{r})$, i.e.,

$$\frac{1}{m}\sum_{i=1}^{m} E_{LL}(s_i \mid \boldsymbol{s}) \approx \frac{1}{n}\sum_{j=1}^{n} E_{LL}(r_j \mid \boldsymbol{r}). \tag{8}$$

Therefore, from the above equation, it could be intuitively seen that the token embeddings in the last layer of the student model could have some degree of alignment effect (if m and n are close to 1). And for the paired sentences of normal length, the word embedding alignment could also be maintained, as shown in the parts (c) and (d) of Fig. 1. However, it is not obvious that part (d) (last layer) has a better alignment effect than part (c) (9th layer). We will show that the last layer is the best choice for cross-lingual word embedding alignment in Sect. 3.4, and denote BERTScore using the last layer embeddings of the student model as metric *BERTScore-MKD*. Nevertheless, the reason why cross-lingual word embedding alignment could be achieved by MKD is still very worthy of in-depth analysis.

3 Experiments

In this section, we evaluate the performance of our metric BERTScore-MKD by correlating its scores 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 evaluation (the input is a source sentence and a system translation sentence): The metric BERTScore-MKD chooses the outputs of the last layer in the model pmmb-v2 as the cross-lingual word embedding function, and takes the $F1$ score (without token importance weighting) in Eq. 3 as its value.

System-level evaluation (the input is a set of source sentences and the corresponding system translation sentences): The mean value of BERTScore-MKD on each pair of the sentences is used as its score for system-level evaluation.

It should be pointed out that the above definitions are for reference-free MT evaluations, and reference-based MT evaluation is implemented by just replacing source sentences with reference sentences.

3.1 Datasets

The source language sentences, and their system and reference translations are collected from the WMT17-19 news translation shared tasks [8,22,23], which contain predictions of 166 translation systems across 16 language pairs in WMT17, 149 translation systems across 14 language pairs in WMT18, and 233 translation systems across 18 language pairs in WMT19. Each language pair in WMT17-19 has about 3,000 source sentences, and each is associated with one reference translation and with the automatic translations generated by participating systems. In this paper, all the into-English language pairs in WMT17-19 are chosen for reference-free MT evaluation.

3.2 Baselines

In this paper, a range of reference-free metrics are chosen to compare with our metric BERTScore-MKD: LASIM and LP [24], UNI and UNI+ [8], YiSi-2 [11], CLP-UMD [15] and SentSim [17]. 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 the multilingual pretrained model XML-R[3] is denoted as BERTScore+XLM-R[4] and is selected to directly compare the cross-lingual word embedding alignment effect with our metric BERTScore-MKD; and reference-based baseline metrics BLEU and sentBLEU [8] are selected as references. It should be pointed out that only the results of the metrics BERTScore-MKD and BERTScore+XLM-R are calculated in this paper, and the results of the other metrics are from their respective papers.

3.3 Results

Evaluation Measures. Pearson correlation (r) and Kendall's Tau correlation (τ) [8] are used as measures for metric evaluations, and are defined as follows:

$$r = \frac{\sum_{i=1}^{n}(H_i - \overline{H})(M_i - \overline{M})}{\sqrt{\sum_{i=1}^{n}(H_i - \overline{H})^2} \cdot \sqrt{\sum_{i=1}^{n}(M_i - \overline{M})^2}}, \tag{9}$$

$$\tau = \frac{|Concordant| - |Discordant|}{|Concordant| + |Discordant|}. \tag{10}$$

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

[4] The 9th layer of XLM-R is chosen for the cross-lingual word embeddings and $F1$ score is used as its metric score according to the recommendations in [3].

Table 1. Segment-level metric results (Pearson correlation) for the into-English language pairs of WMT17. Best results excluding sentBLEU are in bold.

Metrics	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	Avg
sentBLEU	0.435	0.432	0.571	0.404	0.484	0.538	0.512	0.481
SentSim	**0.499**	**0.523**	0.578	0.574	0.551	0.569	**0.600**	0.556
CLP-UMD	0.494	0.462	**0.647**	**0.664**	0.511	0.560	0.528	0.552
BERTScore+XML-R	0.319	0.409	0.414	0.402	0.337	0.382	0.510	0.396
BERTScore-MKD	**0.499**	0.475	0.644	0.584	**0.597**	**0.579**	0.565	**0.563**

Table 2. Segment-level metric results (Kendall's Tau correlation) for the into-English language pairs of WMT19. 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
LASIM	−0.024	-	-	-	-	0.022	-	-
LP	−0.096	-	-	-	-	−0.035	-	-
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
BERTScore+XLM-R	0.084	0.185	0.149	0.176	0.144	0.057	0.157	0.136
BERTScore-MKD	**0.093**	**0.234**	**0.171**	**0.310**	**0.211**	**0.089**	0.208	**0.188**

In Eq. 9, 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. In Eq. 10, *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. It should be pointed out that the measure r could be used for both system-level and segment-level evaluations, while the measure τ is mainly for segment-level evaluation.

Segment-Level Results. Table 1 and Table 2 show the comparison results of the metrics for the reference-free segment-level evaluations on the into-English language pairs of WMT17 and WMT19 respectively.

From the comparison results of BERTScore+XLM-R and BERTScore-MKD in Table 1 and Table 2, it could be seen that BERTScore-MKD has significantly better results on all the into-English language pairs of WMT17 (avg. 0.396 → 0.563) and WMT19 (avg. 0.136 → 0.188), which indicates the cross-lingual word embeddings by MKD have much better alignment effects because only the word embeddings are different for the two metrics.

Table 3. System-level metric results (Pearson correlation) for the into-English language pairs of WMT17. Best results excluding BLEU are in bold.

Metrics	cs-en	de-en	fi-en	lv-en	ru-en	tr-en	zh-en	Avg
BLEU	0.971	0.923	0.903	0.979	0.912	0.976	0.864	0.933
CLP-UMD	**0.984**	0.904	0.861	**0.968**	0.850	0.922	0.817	0.901
BERTScore+XLM-R	0.750	0.692	0.653	0.650	0.332	0.689	0.635	0.629
BERTScore-MKD	0.953	**0.974**	**0.958**	0.871	**0.976**	**0.950**	**0.913**	**0.942**

Table 4. System-level metric results (Pearson correlation) for the into-English language pairs of WMT18. Best results excluding BLEU are in bold.

Metrics	cs-en	de-en	et-en	fi-en	ru-en	tr-en	zh-en	Avg
BLEU	0.970	0.971	0.986	0.973	0.979	0.657	0.978	0.931
CLP-UMD	**0.979**	**0.967**	**0.979**	0.947	0.942	0.673	**0.954**	0.919
BERTScore+XLM-R	−0.528	0.958	0.908	**0.957**	0.905	0.489	0.770	0.637
BERTScore-MKD	0.948	0.963	0.936	0.952	**0.978**	**0.939**	0.925	**0.949**

Table 5. System-level metric results (Pearson correlation) for the into-English language pairs of WMT19. 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
LASIM	0.247	-	-	-	-	0.310	-	-
LP	0.474	-	-	-	-	0.488	-	-
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
CLP-UMD	0.625	0.890	−0.060	**0.993**	0.851	**0.928**	**0.968**	0.742
BERTScore+XLM-R	0.785	0.866	−0.007	0.117	0.657	−0.372	0.728	0.396
BERTScore-MKD	0.823	**0.956**	0.420	0.828	**0.946**	0.747	0.924	**0.806**

And when being compared with the current SOTA metrics involved in this paper, our metric BERTScore-MKD gets the best average scores and ranks first on the all language pairs except zh-en of WMT19 and 3 language pairs (cs-en, ru-en and tr-en) of WMT17. Moreover, as the sentence embeddings of SBERT are adopted in SentSim [17], and BERTScore-MKD uses the word embeddings distilled from SBERT, it could be seen from Table 1 that using word embeddings has better performance than using sentence embeddings (avg. 0.563 *vs.* 0.556), which means using the cross-lingual word embeddings by MKD is a better choice for reference-free MT evaluation.

System-Level Results. Tables 3, 4 and 5 illustrate the comparison results of the metrics for the reference-free system-level evaluations on the into-English language pairs of WMT17, WMT18 and WMT19 respectively.

From the experimental results in Tables 3, 4 and 5, it could be seen again that BERTScore-MKD has significantly better results than BERTScore+XLM-R on all the into-English language pairs of WMT17-19 (avg. 0.629 → 0.942, 0.637 → 0.949, 0.396 → 0.806) except fi-en of WMT18 (0.952 *vs.* 0.957), and gets the best average scores on the into-English language pairs of WMT17-19 when the current SOTA metrics are chosen for comparison. Moreover, the reference-free metric BERTScore-MKD even gets better results than the reference-based metric BLEU on WMT17 and WMT18 (avg. 0.942 *vs.* 0.933, 0.949 *vs.* 0.931).

Therefore, from the segment-level and system-level experimental results in Tables 1, 2, 3, 4 and 5, it could be seen that BERTScore-MKD is very competitive for reference-free MT evaluation when the current SOTA metrics that we know are chosen for comparison. And in Sect. 3.5 we will show that BERTScore-MKD is also suitable for reference-based MT evaluation.

3.4 Effects of Embedding Layers

Since BERTScore is sensitive to the layer of the model selected to generate the contextual token embeddings [3], we investigate which layer of the model pmmb-v2 is the best choice for BERTScore-MKD as a *reference-free* metric through experimental comparisons on the into-English language pairs of WMT19.

BERTScore+XLM-R is chosen for comparison, and the mean values on the into-English language pairs of WMT19 for segment-level and system-level evaluations are illustrated in Fig. 3.

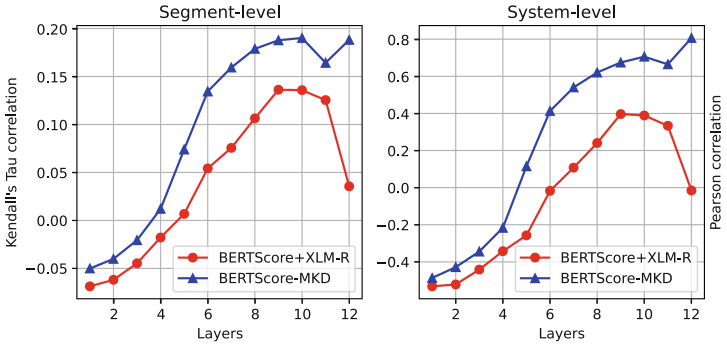

Fig. 3. Mean measure values of BERTScore-MKD and BERTScore+XLM-R with different layers of word embeddings for segment-level and system-level reference-free MT evaluations on the into-English language pairs of WMT19

From Fig. 3, it could be clearly seen that the last layer is the best choice for MKD-BERTScore on both segment-level and system-level evaluations, which

Table 6. System-level reference-based metric results (Pearson correlation) for the into-English language pairs of WMT19. Best results 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
BERTScore+XLM-R	0.932	0.981	**0.919**	**0.998**	**0.992**	0.912	0.962	0.957
BERTScore-MKD^{9th}	0.931	**0.994**	0.897	0.970	0.991	0.971	0.964	**0.960**
BERTScore-MKD^{last}	**0.934**	0.990	0.801	0.943	0.981	**0.974**	**0.968**	0.941

Table 7. System-level reference-based metric results (Pearson correlation) for the from-English language pairs of WMT19. Best results 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
BERTScore+XLM-R	**0.979**	**0.990**	**0.980**	**0.922**	**0.983**	0.978	0.985	**0.929**	**0.968**
BERTScore-MKD^{9th}	0.966	0.986	0.956	0.899	0.980	0.938	**0.991**	0.871	0.948
BERTScore-MKD^{last}	0.942	0.982	0.928	0.889	0.972	0.876	0.985	0.814	0.924

is consistent with our 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, it could be also found that our metric BERTScore-MKD outperforms BERTScore+XLM-R on every layer for both segment-level and system-level reference-free MT evaluations.

3.5 As Reference-Based Metric

In this section we investigate the performance of BERTScore-MKD as a reference-based metric, where source sentences in the input are replaced with reference sentences. As the system translations and the reference sentences are in the same language, there is no need for cross-lingual alignment. Therefore, besides the last layer, BERTScore-MKD also uses the outputs of the 9th layers (recommended in [3]) in the model pmmb-v2 as the contextual word embedding function.

Table 6 and Table 7 report the results of BERTScore-MKD as a reference-base metric for system-level evaluations on the into-English and from-English language pairs of WMT19, and the metrics BLEU and BERTScore+XLM-R are chosen for comparison.

From the comparison results in Table 6 and Table 7, it could be seen that both BERTScore+XLM-R and BERTScore-MKD are clearly better than the classical metric BLEU, and our metric BERTScore-MKD is almost the same with the current SOTA metric BERTScore+XLM-R. Meanwhile, the 9th layer is slightly better than the last layer for BERTScore-MKD. In summary, BERTScore-MKD shows its effectiveness and robustness as a reference-base metric.

4 Conclusion

In this paper, it is found out that the cross-lingual word embedding alignment could be achieved implicitly through multilingual knowledge distillation (MKD) for sentence embedding alignment. With the framework of BERTScore, a reference-free metric BERTScore-MKD is proposed by incorporating MKD into MT evaluation. As shown in the performance test of BERTScore-MKD on the into-English language pairs of WMT17-19 for both segment-level and system level evaluations, the reference-free metric BERTScore-MKD is very competitive (best mean scores on WMT17-19 and better than BLEU on WMT17-18) with the current SOTA metrics that we know. Furthermore, the comparison results on WMT19 show the effectiveness and robustness of BERTScore-MKD as a reference-base metric. Although we have found that MKD could achieve the alignment of cross-lingual word embeddings and the last layer of the distilled student model is the best choice for reference-free MT evaluation, the reason why MKD could achieve the alignment is still worthy of further study.

References

1. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pp. 311–318. Association for Computational Linguistics, July 2002
2. Lavie, A., Agarwal, A.: METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic, pp. 228–231. Association for Computational Linguistics, June 2007
3. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: evaluating text generation with BERT. In: 8th International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020. OpenReview.net (2020)
4. Sellam, T., Das, D., Parikh, A.: BLEURT: learning robust metrics for text generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7881–7892, July 2020
5. Zaidan, O.F., Callison-Burch, C.: Crowdsourcing translation: professional quality from non-professionals. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, pp. 1220–1229. Association for Computational Linguistics, June 2011
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics, June 2019
7. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. In: ACL 2020, 5–10 July 2020, pp. 8440–8451. Association for Computational Linguistics (2020)

8. Ma, Q., Wei, J., Bojar, O., Graham, Y.: 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, Florence, Italy. Shared Task Papers, vol. 2, pp. 62–90. Association for Computational Linguistics, August 2019
9. Popović, M., Vilar, D., Avramidis, E., Burchardt, A.: Evaluation without references: IBM1 scores as evaluation metrics. In: Proceedings of the Sixth Workshop on Statistical Machine Translation, Edinburgh, Scotland, pp. 99–103, July 2011
10. Specia, L., Shah, K., de Souza, J.G., Cohn, T.: QuEst - a translation quality estimation framework. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Sofia, Bulgaria, pp. 79–84. Association for Computational Linguistics, August 2013
11. Lo, C.k.: 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), Florence, Italy, pp. 507–513. Association for Computational Linguistics, August 2019
12. Thompson, B., Post, M.: 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), pp. 90–121. Association for Computational Linguistics, November 2020
13. Rei, R., et al.: Are references really needed? Unbabel-IST 2021 submission for the metrics shared task. In: Proceedings of the Sixth Conference on Machine Translation, pp. 1030–1040, November 2021
14. Rei, R., Stewart, C., Farinha, A.C., Lavie, A.: COMET: a neural framework for MT evaluation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2685–2702, November 2020
15. Zhao, W., Glavaš, G., Peyrard, M., Gao, Y., West, R., Eger, S.: On the limitations of cross-lingual encoders as exposed by reference-free machine translation evaluation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1656–1671, July 2020
16. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
17. Song, Y., Zhao, J., Specia, L.: SentSim: crosslingual semantic evaluation of machine translation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3143–3156, June 2021
18. Wan, Y., et al.: UniTE: unified translation evaluation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, pp. 8117–8127. Association for Computational Linguistics, May 2022
19. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 4512–4525. Association for Computational Linguistics, November 2020
20. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Hong Kong, China, pp. 3982–3992. Association for Computational Linguistics, November 2019
21. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)

22. Bojar, O., Graham, Y., Kamran, A.: Results of the WMT17 metrics shared task. In: Proceedings of the Second Conference on Machine Translation, Copenhagen, Denmark, pp. 489–513. Association for Computational Linguistics, September 2017
23. Ma, Q., Bojar, O., Graham, Y.: Results of the WMT18 metrics shared task: both characters and embeddings achieve good performance. In: Proceedings of the Third Conference on Machine Translation: Shared Task Papers, Belgium, Brussels, pp. 671–688. Association for Computational Linguistics, October 2018
24. Yankovskaya, E., Tättar, A., Fishel, M.: 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), Florence, Italy, pp. 101–105. Association for Computational Linguistics, August 2019