Incorporating Multilingual Knowledge Distillation into Machine Translation Evaluation

CCKS2022

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

Introduction

- 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).

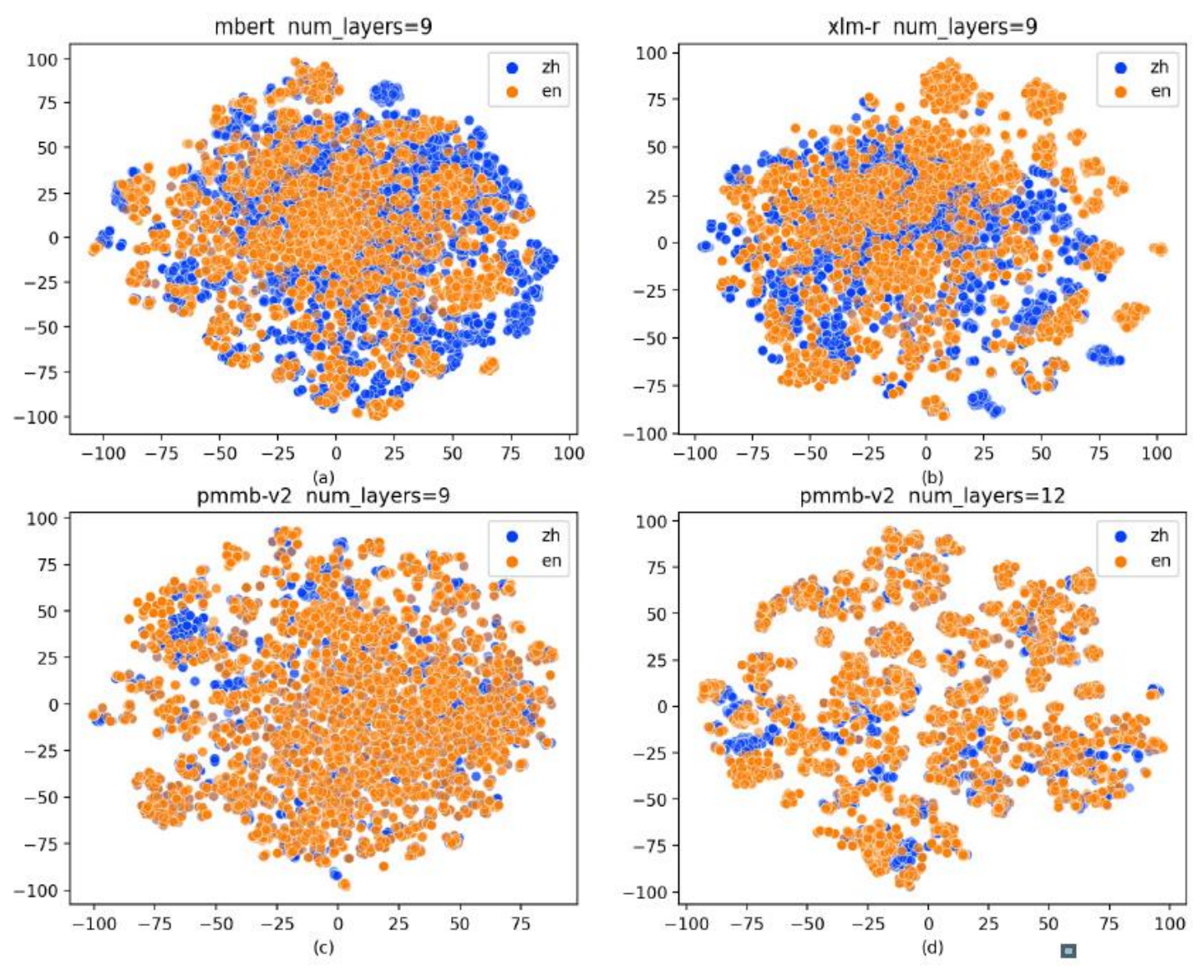


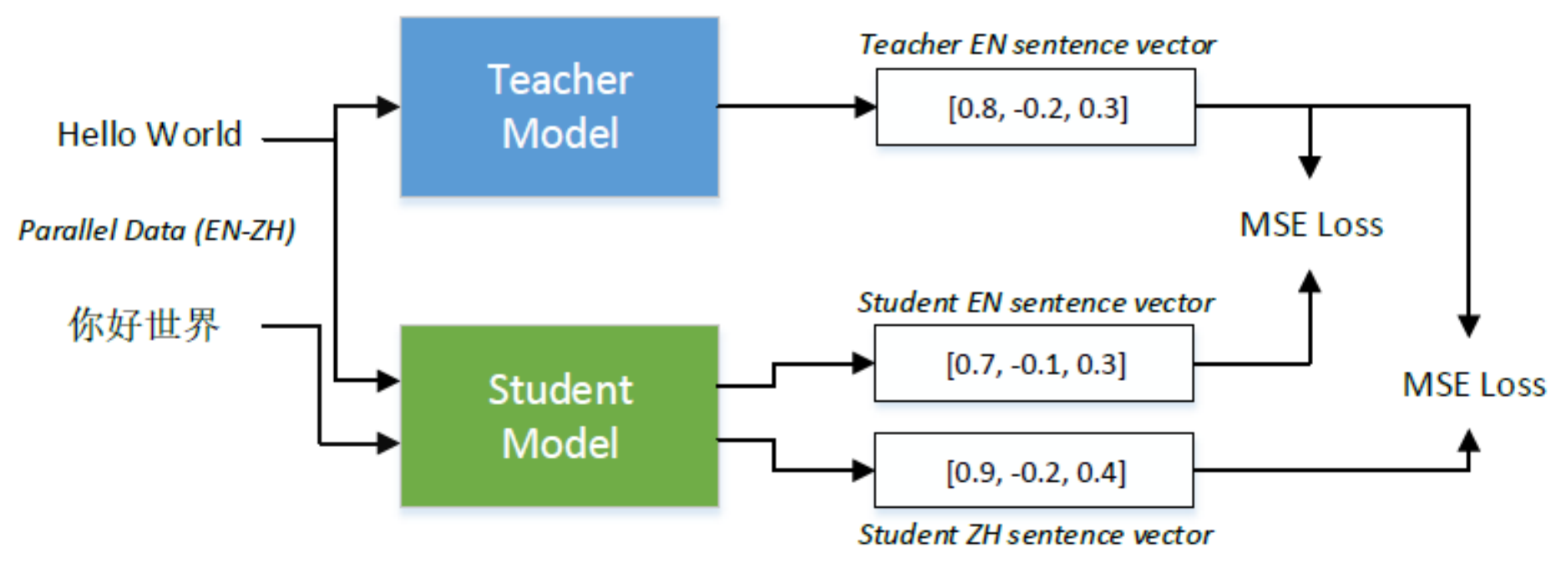
Fig. 1. First two principle components of contextual token embeddings of mBERT, XLM-R and pmmb-v2 for 100 zh-en parallel sentences in WMT19 by t-SNE (The more areas that do not cover each other, the worse the word embedding alignment effectiveness)

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

- With the framework of BERTScore, we propose a metric BERTScore-MKD for reference-free machine translation evaluation.

Methods

1. Multilingual Knowledge Distillation



2. BERTScore-MKD

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

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

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

Results

- Segment-level results

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**

- System-level results

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**

Discussion

- Effects of Embedding Layers

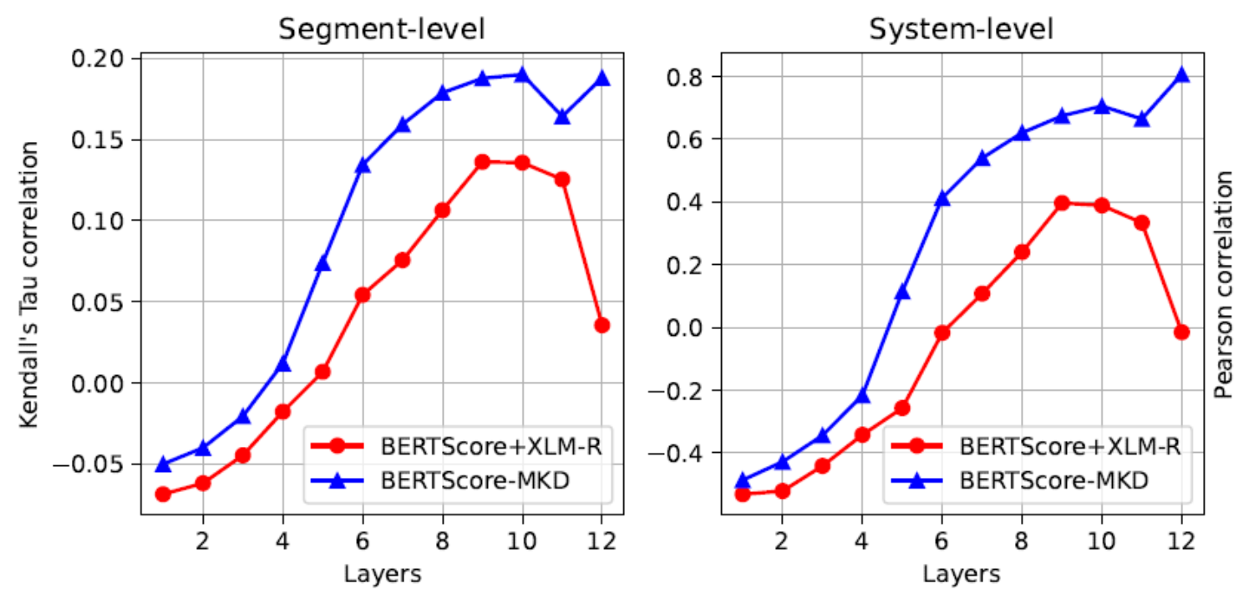


Fig. 3. Mean measure values of BERTScore-MKD and BERTScore+XLM-R with different layers of word embeddings for segment-level and system-level reference-free MT evaluations on the into-English language pairs of WMT19

Supplementary

- As Reference-based Metric

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

华为翻译中心
zhangmin186@huawei.com

2022全国知识图谱与语义计算大会
China Conference on Knowledge Graph and Semantic Computing
秦皇岛 8.24-8.27