Target-Side Language Model for Reference-Free Machine Translation Evaluation

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

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

Abstract. With the rapid progress of deep learning in multilingual language processing, there has been a growing interest in reference-free machine translation evaluation, where source texts are directly compared with system translations. In this paper, we design a reference-free metric that is based only on a target-side language model for segment-level and system-level machine translation evaluations respectively, and it is found out that promising results could be achieved when only the target-side language model is used in such evaluations. From the experimental results on all the 18 language pairs of the WMT19 news translation shared task, it is interesting to see that the designed metrics with the multilingual model XLM-R get very promising results (best segment-level mean score on the from-English language pairs, and best system-level mean scores on the from-English and none-English language pairs) when the current SOTA metrics that we know are chosen for comparison.

Keywords: Target-side language model · Machine translation evaluation · Reference-free metric

1 Introduction

Traditional automatic metrics for machine translation (MT) score MT output by comparing it with one or more reference translations. Common such metrics 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 reference translations [5].

These problems can be avoided through *reference-free* MT evaluation, meaning that only source texts are used in MT output evaluation and they are directly compared with system translations. Recently, with the rapid progress of deep learning in multilingual language processing [6,7], a lot of reference-free metrics have been proposed for such evaluation. Popović et al. [8] exploited a bag-of-word translation model for quality estimation, which sums over the likelihoods

T. Xiao and J. Pino (Eds.): CCMT 2022, CCIS 1671, pp. 45–53, 2022.
https://doi.org/10.1007/978-981-19-7960-6_5

of aligned word pairs between source and translation texts. Specia et al. [9] used language-agnostic linguistic features extracted from source texts and system translations to estimate quality. YiSi-2 [10] evaluates system translations by summing similarity scores over words pairs which are best-aligned mutual translations. Moreover, by introducing cross-lingual linear projection, Lo and Larkin [11] greatly improved the effect of YiSi-2. 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. Gekhman et al. [15] proposed a simple and effective Knowledge-Based Evaluation (KoBE) method by measuring the recall of entities found in source texts and system translations. To mitigate the misalignment of cross-lingual word embedding spaces, Zhao et al. [16] proposed post-hoc re-alignment strategies which integrate a target-side GPT [17] language model. Song et al. [18] proposed an unsupervised metric SentSim by incorporating a notion of sentence semantic similarity.

In this paper, we find out that assessing system translation only with a target-side language model could achieve very promising results. With a modified sentence perplexity calculation for system translations, we design a reference-free metric for segment-level and system-level MT evaluations respectively. And then we test the performances of the two metrics on all the 18 language pairs of WMT19 news translation shared task [19]. The experimental results demonstrate that our metrics with the pretrained model XLM-R [7] are very competitive for reference-free MT evaluations when compared with the current SOTA reference-free metrics that we know.

2 Target-Side Language Model Metrics

A statistical language model is a probability distribution over sequences of words [20]. Given such a sequence with m words, i.e., $\boldsymbol{s} = (w_1, \ldots, w_m)$, it assigns a probability $P(\boldsymbol{s})$ to the whole sequence, which is defined as:

$$P(\boldsymbol{s}) = P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i | w_1, \ldots, w_{i-1}). \tag{1}$$

In order to overcome the data sparsity problem in building a statistical language model, a common solution is to assume that the probability of a word only depends on the previous n words. This is known as the n-gram model or unigram model when $n = 1$. So the probability $P(\boldsymbol{s})$ could be approximated as:

$$P(\boldsymbol{s}) = \prod_{i=1}^{m} P(w_i | w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i | w_{i-(n-1)}, \ldots, w_{i-1}). \tag{2}$$

With the advancements in deep learning [21], various neural language models are proposed to use continuous representations or embeddings of words to make

their predictions [6,22]. Typically, a neural language model is constructed and trained as probabilistic classifiers for

$$P(w \mid context), \ for \ w \in V. \tag{3}$$

That is to say, the model is trained to predict a probability distribution over the vocabulary V, when some linguistic *context* is given.

In this paper, we adopt the masked language model [6] to design a reference-free metric for segment-level and system-level MT evaluations respectively.

For segment-level evaluation where a single system translation sentence $\boldsymbol{s}$ is provided, the metric SEG_LM is defined as:

$$SEG\_LM(\boldsymbol{s}) = \frac{1}{m}\sum_{i=1}^{m} \log \frac{1}{P(w_i|\boldsymbol{s}-w_i)}, \tag{4}$$

where m is the number of words in sentence $\boldsymbol{s}$, w_i is the i-th word in $\boldsymbol{s}$, and $P(w_i|\boldsymbol{s}-w_i)$ the probability of w_i predicted by the masked language model when w_i is replaced by `[MASK]` in $\boldsymbol{s}$.

It should be pointed out that the metric SEG_LM is slightly different from the log form of the sentence perplexity [20] (PPL), which is defined as:

$$\log PPL(\boldsymbol{s}) = \log \sqrt[m]{\frac{1}{P(w_1,\ldots,w_m)}} = \frac{1}{m}\sum_{i=1}^{m} \log \frac{1}{P(w_i|w_1,\ldots,w_{i-1})}. \tag{5}$$

From the above definitions, it could be seen that the context for predicting the probability of w_i in PPL is different from SEG_LM.

For system-level evaluation where a set of system translation sentences S is given, the metric SYS_LM is defined as:

$$SYS\_LM(S) = \frac{1}{|S|}\sum_{\boldsymbol{s}\in S} SEG\_LM(\boldsymbol{s}), \tag{6}$$

which is the mean value of SEG_LM scores on each sentence in S.

Although source texts are not considered in our designed metrics, the experimental results on WMT19 in Sect. 3 will show that the metrics SEG_LM and SYS_LM are very promising for both segment-level and system-level reference-free MT evaluations.

3 Experiments

In this section, we evaluate the performance of our metrics SEG_LM and SYS_LM by correlating their scores with human judgments of translation quality for reference-free MT evaluations. The pretrained multilingual model XLM-R[1] is used as the masked language model for our metrics.

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

Table 1. Segment-level metric results for the into-English language pairs of WMT19

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
YiSi-2+CLP	**0.116**	**0.271**	**0.249**	**0.370**	**0.281**	**0.121**	**0.340**	**0.250**
SEG_LM	0.115	0.265	0.214	0.135	0.280	0.120	0.183	0.187

Table 2. Segment-level metric results for the from-English language pairs of WMT19

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
LASIM	–	0.147	–	–	–	–	−0.240	–	–
LP	–	−0.119	–	–	–	–	−0.158	–	–
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
SEG_LM	**0.443**	**0.343**	**0.492**	0.328	0.301	**0.471**	**0.457**	**0.297**	**0.392**

Table 3. Segment-level metric results for the none-English language pairs of WMT19

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**
SEG_LM	0.263	0.244	0.198	0.235

3.1 Datasets and Baselines

The source language sentences, and their system and reference translations are collected from the WMT19 news translation shared task [19], which contains predictions of 233 translation systems across 18 language pairs. Each language pair has about 3,000 source sentences, and each is associated with one reference translation and with the automatic translations generated by the participating systems. In this paper, all the 18 language pairs in WMT19 are chosen for reference-free MT evaluation.

A range of reference-free metrics are chosen to compare with our metrics: LASIM and LP [23], UNI and UNI+ [19], YiSi-2 [10] and YiSi-2+CLP [11], KoBE [15] and CLP-UMD [16]. To the best of our knowledge, the above metrics could cover most of the current SOTA metrics for reference-free MT evaluation. Reference-based baseline metrics BLEU and sentBLEU [24] are selected as references. It should be pointed out that only the results of our metrics SEG_LM and SYS_LM are calculated in this paper, and the results of the other metrics are from their respective papers.

3.2 Results

Evaluation Measures. Kendall's Tau and Pearson correlations [19] are used as measures for segment-level and system-level metric evaluations respectively.

Table 4. System-level metric results for the into-English language pairs of WMT19

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
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
CLP-UMD	0.625	0.890	−0.060	**0.993**	0.851	**0.928**	0.968	0.742
SYS_LM	0.856	0.932	0.748	0.696	0.932	0.869	0.480	0.788

Segment-level Results. Tables 1, 2 and 3 show the comparison results of the metrics for reference-free segment-level evaluations on the into-English, from-English and none-English language pairs of WMT19 respectively (Best results excluding sentBLEU are in bold).

From Table 1, it could be seen that the scores of our metric SEG_LM on the de-en, lt-en and ru-en language pairs are very close to the best values (only 0.001 gap). And as shown in Table 2, our metric not only gets the best mean score on the from-English language pairs, but also ranks first on 6 of all the 8 language pairs. The results in Table 3 show that our metric even gets better scores on all the none-English language pairs than the reference-based metric sentBLEU. Therefore, our metric SEG_LM is very promising for segment-level MT evaluation especially when the target-side language is not English.

Table 5. System-level metric results for the from-English language pairs of WMT19

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
LASIM	–	0.871	–	–	–	–	0.823	–	–
LP	–	0.569	–	–	–	–	−0.661	–	–
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
SYS_LM	**0.896**	**0.978**	**0.941**	0.683	0.897	**0.919**	0.819	**0.959**	**0.886**

Table 6. System-level metric results for the none-English language pairs of WMT19

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
KoBE	**0.958**	0.485	−0.785	0.219
SYS_LM	0.885	**0.902**	**0.778**	**0.855**

System-level Results. Tables 4, 5 and 6 illustrate the comparison results of the metrics for reference-free system-level evaluations on the into-English, from-English and none-English language pairs of WMT19 respectively (Best results excluding BLEU are in bold).

As shown in the into-English results of Table 4, our metric SYS_LM again gets scores very close to the best values on the fi-en and lt-en language pairs. The results in Table 5 demonstrate that our metric gets the best mean score and 5 best scores on all the 8 from-English language pairs. Meanwhile, the results in Table 6 show that SYS_LM gets better scores than the SOTA metric YiSi-2+CLP on the system-level evaluations, although it does not outperform YiSi-2+CLP on the segment-level evaluations, as shown in Table 3. In addition, SYS_LM gets the best mean score on the none-English language pairs. Overall, the experimental results demonstrate that our metric SYS_LM is very competitive for system-level MT evaluations when the current SOTA metrics that we know are involved for comparison.

3.3 Discussion

In this section, an explanation for why target-side language model works is provided. For segment-level evaluation where the input is a source sentence $\boldsymbol{s}$ and a

system translation sentence $\boldsymbol{t}$, we design metrics to estimate the true probability $P(\boldsymbol{t}|\boldsymbol{s})$. According to the conditional probability formula, we could have:

$$\log P(\boldsymbol{t}|\boldsymbol{s}) = \log \frac{P(\boldsymbol{s}|\boldsymbol{t})P(\boldsymbol{t})}{P(\boldsymbol{s})} = \log P(\boldsymbol{s}|\boldsymbol{t}) + \log P(\boldsymbol{t}) - \log P(\boldsymbol{s}). \tag{7}$$

The target-side language model is mainly to approximate the second term $\log P(\boldsymbol{t})$, and when there are no much differences in the first term $\log P(\boldsymbol{s}|\boldsymbol{t})$, our target-side language model metric works for MT evaluation.

4 Conclusion

In this paper, a reference-free metric designed only with a target-side language model is proposed for segment-level and system-level MT evaluations respectively. With the pretrained multilingual model XLM-R as the target-side language model, the performances of our metrics SEG_LM and SYS_LM are evaluated on all the 18 language pairs of WMT19. The experimental results show that our metrics are very competitive (best mean score of segment-level evaluations on the from-English language pairs, and best mean scores of system-level evaluations on the from-English and none-English language pairs) when most of the current SOTA reference-free metrics are chosen for comparison. Furthermore, the reason why the target-side language model works is discussed. The fusion of our metrics and other metrics that are for the first term $\log P(\boldsymbol{s}|\boldsymbol{t})$ in Eq. 7 will be our future work.

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, pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (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, pp. 228–231. Association for Computational Linguistics, Prague, Czech Republic (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, ICLR 2020, 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. Association for Computational Linguistics, Online (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, pp. 1220–1229. Association for Computational Linguistics, Portland, Oregon, USA (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, Vol. 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019)
7. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, pp. 8440–8451, 5–10 July 2020. Association for Computational Linguistics (2020)
8. 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, pp. 99–103. Association for Computational Linguistics, Edinburgh, Scotland (2011)
9. 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, pp. 79–84. Association for Computational Linguistics, Sofia, Bulgaria (2013)
10. 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), pp. 507–513. Association for Computational Linguistics, Florence, Italy (2019)
11. Lo, C.K., Larkin, S.: Machine translation reference-less evaluation using YiSi-2 with bilingual mappings of massive multilingual language model. In: Proceedings of the Fifth Conference on Machine Translation, pp. 903–910. Association for Computational Linguistics, Online (2020)
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, Online (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 (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 (2020)
15. Gekhman, Z., Aharoni, R., Beryozkin, G., Freitag, M., Macherey, W.: KoBE: knowledge-based machine translation evaluation. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 3200–3207. Association for Computational Linguistics (2020)
16. 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. Association for Computational Linguistics (2020)
17. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
18. 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. Association for Computational Linguistics (2021)
19. 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 (Volume 2: Shared Task Papers, Day 1), pp. 62–90. Association for Computational Linguistics, Florence, Italy (2019)
20. Rosenfeld, R.: Two decades of statistical language modeling: where do we go from here. In: Proceedings of the IEEE, vol. 88, pp. 1270–1278 (2000)
21. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science **313**(5786), 504–507 (2006)
22. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013)
23. 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), pp. 101–105. Association for Computational Linguistics, Florence, Italy (2019)
24. Koehn, P., et al.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pp. 177–180. Association for Computational Linguistics, Prague, Czech Republic (2007)