Evaluating Machine Translation without Human References Using Cross-lingual Encoders

Wei Zhao[†], **Yang Gao**[Φ], **Steffen Eger**[†]
[†] Computer Science Department, Technische Universität Darmstadt, Germany
[Φ] Computer Science Department, Royal Holloway, University of London, UK
zhao@aiphes.tu-darmstadt.de, yang.gao@rhul.ac.uk
eger@aiphes.tu-darmstadt.de

In machine translation (MT), the classical approach to evaluation is to compare system translation y' against reference y using a metric $\mathfrak{m}$:

$$\mathfrak{m}(y, y')$$

Historically, the metric $\mathfrak{m}$ was (a variant of) the BLEU score, which strictly counts the exact n-gram matches between y and y'. Our recent work (Zhao et al., 2019) has shown that a soft-matching metric named MoverScore, combining contextualized word embeddings and Earth Mover Distance (Rubner et al., 2000), correlates substantially better with human assessments of translation quality than BLEU, as it allows for lexical variation in y and y'. This is a fundamental step towards better MT evaluation.

While our proposed metric still requires parallel data (y,y'), metrics that instead use source sentence x and system translation y' as arguments:

$$\mathfrak{m}(x, y')$$

would be more desirable, as they (a) do not require references to assess the quality of systems, (b) therefore allow evaluation in any domain.

In our ongoing work, we investigate reference-free metrics with soft-matching approaches to compare MT systems. One obvious solution is to encode x and y' using *cross-lingual* embeddings and measure the distance between them with MoverScore or cosine similarity.

The intriguing finding so far (see Table 1) is that metrics $\mathfrak{m}(x, y')$ built from cross-lingual representations perform substantially below lexical-matching metrics $\mathfrak{m}(y, y')$ like BLEU, while embedding based metrics computed on y and y' can clearly outperform the latter. We speculate that two reasons may explain our findings:

(i) Mismatch between evaluation scores $\mathfrak{m}(x, y')$ and human ratings, which are instead based on y and y'.

(ii) Current cross-lingual embeddings do not adequately capture cross-lingual similarity between source sentences and system translations.

Our experiments We obtain source sentences, system and reference translations from the WMT 2017 news translation shared task (Bojar et al., 2017). The compared metrics are: SentBLEU (Koehn et al., 2007), LASER (Artetxe and Schwenk, 2018), XLM (Lample and Conneau, 2019), MoverScore (Zhao et al., 2019) and XMoverScore[1].

Setting	Metrics	Type	de-en	ru-en	zh-en
$\mathfrak{m}(y, y')$	SentBLEU	X	0.432	0.484	0.511
	MoverScore	♠	0.708	0.738	0.744
$\mathfrak{m}(x, y')$	XLM	♡	0.330	0.279	0.259
	LASER	♡	0.402	0.363	0.466
	XMoverScore	♠	0.343	0.222	0.360

Table 1: Pearson correlations with segment-level human judgments. ♠ and ♡ denote cross-lingual word and sentence embeddings, respectively. For sentence embeddings, we compute the cosine similarity between the embeddings of x and y' as metric score $\mathfrak{m}(x, y')$ and for word embeddings we use MoverScore as metric, which integrates contextualized (cross- or monolingual word embeddings) into unigram-based Word Mover Distance.

In Table 1, we observe that the best correlation is achieved by our embedding-based metric MoverScore, computed from y and y'. Interestingly, no $\mathfrak{m}(x, y')$ metric produces an even moderate correlation with human judgments; in fact, XMoverScore's correlation with humans is about half of the correlation of MoverScore. We will investigate modifications to current cross-lingual embeddings for the little explored task of cross-lingual metric induction from system translations

[1] XMoverScore is a cross-lingual variant of MoverScore without finetuning on MNLI.

against source sentences.

References

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *CoRR*, abs/1812.10464.

Ondrej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Conference on Machine Translation (WMT)*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. 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*, pages 177–180.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China.