| cases | doc_1 | doc_2 | decision | id |
|---|---|---|---|---|
| | **authors**: • Wei Zhao • Goran GlavaÅ¡ • Maxime Peyrard • Yang Gao • Robert West • Steffen Eger<br>**title**: On the Limitations of Cross-lingual Encoders as Exposed by Reference-Free Machine Translation Evaluation<br>**publication_date**: 2020-05-01 00:00:00<br>**source**: SupportedSources.OPENALEX<br>**journal**: Meeting of the Association for Computational Linguistics<br>**volume**:<br>**doi**: 10.18653/v1/2020.acl-main.151<br>**urls**: • https://openalex.org/W3035459196 • https://doi.org/10.18653/v1/2020.acl-main.151 • https://www.aclweb.org/anthology/2020.acl-main.151.pdf<br>**id**: id-4849408777373638810<br>**abstract**:<br>**versions**: | **authors**: • Steffen Eger • Goran GlavaÅ¡ • Yang Gao • Robert West • Wei Zhao • Maxime Peyrard<br>**title**: On the Limitations of Cross-lingual Encoders as Exposed by Reference-Free Machine Translation Evaluation<br>**publication_date**: 2020-05-03 00:00:00<br>**source**: SupportedSources.PAPERS_WITH_CODE<br>**journal**:<br>**volume**:<br>**doi**:<br>**urls**: • https://arxiv.org/pdf/2005.01196v3.pdf • https://github.com/AIPHES/ACL20-Reference-Free-MT-Evaluation • https://aclanthology.org/2020.acl-main.151.pdf<br>**id**: id8263332987419513617<br>**abstract**: Evaluation of cross-lingual encoders is usually performed either via zero-shot cross-lingual transfer in supervised downstream tasks or via unsupervised cross-lingual textual similarity. In this paper, we concern ourselves with reference-free machine translation (MT) evaluation where we directly compare source texts to (sometimes low-quality) system translations, which represents a natural adversarial setup for multilingual encoders. Reference-free evaluation holds the promise of web-scale comparison of MT systems. We systematically investigate a range of metrics based on state-of-the-art cross-lingual semantic representations obtained with pretrained M-BERT and LASER. We find that they perform poorly as semantic encoders for reference-free MT evaluation and identify their two key limitations, namely, (a) a semantic mismatch between representations of mutual translations and, more prominently, (b) the inability to punish "translationese", i.e., low-quality literal translations. We propose two partial remedies: (1) post-hoc re-alignment of the vector spaces and (2) coupling of semantic-similarity based metrics with target-side language modeling. In segment-level MT evaluation, our best metric surpasses reference-based BLEU by 5.7 correlation points.<br>**versions**: | DUPLICATES | 254 |