		doc_1	doc_2	decision	id
cases	authors	Wei Zhao Goran GlavaÅ¡ Maxime Peyrard Yang Gao Robert West Steffen Eger	Wei Zhao Goran GlavaÅ¡ Maxime Peyrard Yang Gao Robert West Steffen Eger	NOT DUPLICATES	414
	title	On the Limitations of Cross-lingual Encoders as Exposed by Reference-Free Machine Translation Evaluation	On the Limitations of Cross-lingual Encoders as Exposed by Reference-Free Machine Translation Evaluation		
	publication_date	2020-05-03 00:00:00	2020-06-08 00:00:00		
	source	SupportedSources.OPENALEX	SupportedSources.INTERNET_ARCHIVE		
	journal	arXiv (Cornell University)			
	volume				
	doi	10.48550/arxiv.2005.01196			
	urls	https://openalex.org/W3020922140 https://doi.org/10.48550/arxiv.2005.01196 http://arxiv.org/pdf/2005.01196	https://web.archive.org/web/20200907173706/https://arxiv.org/pdf/2005.01196v3.pdf		
	id	id-4321866379358349889	id-2115329177077088540		
	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.		
	versions				