

cases	doc_1		doc_2		decision	id
					NOT DUPLICATES	413
	authors	<ul style="list-style-type: none">Wei ZhaoGoran GlavaÅ¡Maxime PeyrardYang GaoRobert WestSteffen Eger	authors	<ul style="list-style-type: none">Wei ZhaoGoran GlavaÅ¡Maxime PeyrardYang GaoRobert WestSteffen Eger		
		title		On the Limitations of Cross-lingual Encoders as Exposed by Reference-Free Machine Translation Evaluation		
	publication_date	2020-05-03 00:00:00	publication_date	2020-05-03 22:10:23+00:00		
	source	SupportedSources.OPENALEX	source	SupportedSources.ARXIV		
	journal	arXiv (Cornell University)	journal	None		
	volume		volume			
	doi	10.48550/arxiv.2005.01196	doi			
	urls	<ul style="list-style-type: none">https://openalex.org/W3020922140https://doi.org/10.48550/arxiv.2005.01196http://arxiv.org/pdf/2005.01196	urls	<ul style="list-style-type: none">http://arxiv.org/pdf/2005.01196v3http://arxiv.org/abs/2005.01196v3http://arxiv.org/pdf/2005.01196v3		
	id	id-4321866379358349889	id	id-3744035111753124581		
	abstract		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			versions		