cases	doc_1		doc_2		decision	id
	authors	• Wei Zhao • Maxime Peyrard • Fei Liu • Yang Gao • Christian M. Meyer, Steffen Eger	**authors**	• Wei Zhao • Maxime Peyrard • Fei Liu • Yang Gao • Christian M. Meyer • Steffen Eger	DUPLICATES	347
	title	MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance	**title**	MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance		
	publication_date	2019-09-26 00:00:00	**publication_date**	2019-01-01 00:00:00		
	source	SupportedSources.INTERNET_ARCHIVE	**source**	SupportedSources.INTERNET_ARCHIVE		
	journal		**journal**	Association for Computational Linguistics		
	volume		**volume**			
	doi		**doi**	10.18653/v1/d19-1053		
	urls	• https://web.archive.org/web/20200930133613/https://arxiv.org/pdf/1909.02622v2.pdf	**urls**	• https://web.archive.org/web/20191203105116/https://www.aclweb.org/anthology/D19-1053.pdf		
	id	id995208829855123467	**id**	id-6049473829661719424		
	abstract	A robust evaluation metric has a profound impact on the development of text generation systems. A desirable metric compares system output against references based on their semantics rather than surface forms. In this paper we investigate strategies to encode system and reference texts to devise a metric that shows a high correlation with human judgment of text quality. We validate our new metric, namely MoverScore, on a number of text generation tasks including summarization, machine translation, image captioning, and data-to-text generation, where the outputs are produced by a variety of neural and non-neural systems. Our findings suggest that metrics combining contextualized representations with a distance measure perform the best. Such metrics also demonstrate strong generalization capability across tasks. For ease-of-use we make our metrics available as web service.	**abstract**	A robust evaluation metric has a profound impact on the development of text generation systems. A desirable metric compares system output against references based on their semantics rather than surface forms. In this paper we investigate strategies to encode system and reference texts to devise a metric that shows a high correlation with human judgment of text quality. We validate our new metric, namely MoverScore, on a number of text generation tasks including summarization, machine translation, image captioning, and data-to-text generation, where the outputs are produced by a variety of neural and non-neural systems. Our findings suggest that metrics combining contextualized representations with a distance measure perform the best. Such metrics also demonstrate strong generalization capability across tasks. For ease-of-use we make our metrics available as web service. 1		
	versions		**versions**			