	doc_1		doc_2		decision	id
cases		authors	• Qingyu Chen • Jingcheng Du • Sun Kim • W. John Wilbur • Zhiyong Lu			
		title	Deep learning with sentence embeddings pre-trained on biomedical corpora improves the performance of finding similar sentences in electronic medical records			
		publication_date	2019-09-06 17:56:01+00:00			
		source	SupportedSources.ARXIV			
		journal	None			
		volume				
		doi				
		urls	• http://arxiv.org/pdf/1909.03044v1 • http://arxiv.org/abs/1909.03044v1 • http://arxiv.org/pdf/1909.03044v1			
		id	id-1373466988577229111			
		abstract	Capturing sentence semantics plays a vital role in a range of text mining applications. Despite continuous efforts on the development of related datasets and models in the general domain, both datasets and models are limited in biomedical and clinical domains. The BioCreative/OHNLP organizers have made the first attempt to annotate 1,068 sentence pairs from clinical notes and have called for a community effort to tackle the Semantic Textual Similarity (BioCreative/OHNLP STS) challenge. We developed models using traditional machine learning and deep learning approaches. For the post challenge, we focus on two models: the Random Forest and the Encoder Network. We applied sentence embeddings pre-trained on PubMed abstracts and MIMIC-III clinical notes and updated the Random Forest and the Encoder Network accordingly. The official results demonstrated our best submission was the ensemble of eight models. It achieved a Person correlation coefficient of 0.8328, the highest performance among 13 submissions from 4 teams. For the post challenge, the performance of both Random Forest and the Encoder Network was improved; in particular, the correlation of the Encoder Network was improved by ~13%. During the challenge task, no end-to-end deep learning models had better performance than machine learning models that take manually-crafted features. In contrast, with the sentence embeddings pre-trained on biomedical corpora, the Encoder Network now achieves a correlation of ~0.84, which is higher than the original best model. The ensembled model taking the improved versions of the Random Forest and Encoder Network as inputs further increased performance to 0.8528. Deep learning models with sentence embeddings pre-trained on biomedical corpora achieve the highest performance on the test set.			
		versions				
		authors	• Qingyu Chen • Jingcheng Du • Sun Kim • W. Wilbur • Zhiyong Lu		DUPLICATES	346
		title	Deep learning with sentence embeddings pre-trained on biomedical corpora improves the performance of finding similar sentences in electronic medical records			
		publication_date	2019-09-06 00:00:00			
		source	SupportedSources.SEMANTIC_SCHOLAR			
		journal	BMC Medical Informatics and Decision Making			
		volume	20			
		doi	10.1186/s12911-020-1044-0			
		urls	• https://www.semanticscholar.org/paper/16dc153aaeef8f8e62bd7010f9ab2032ba671a11			
		id	id2536516996323524334			
		abstract	None			
		versions				