		doc_1		doc_2		decision	id
cases	authors	• Wang, Y. • Afzal, N. • Fu, S. • Wang, L. • Shen, F. • Rastegar-Mojarad, M. • Liu, H.	authors	• Yanshan Wang • Naveed Afzal • Sunyang Fu • Liwei Wang • Feichen Shen • Majid Rastegar-Mojarad • Hongfang Liu		DUPLICATES	360
	title	MedSTS: a resource for clinical semantic textual similarity	title	MedSTS: A Resource for Clinical Semantic Textual Similarity			
	publication_date	2018-10-24 00:00:00	publication_date	2018-08-28 16:43:19+00:00			
	source	SupportedSources.CROSSREF	source	SupportedSources.ARXIV			
	journal		journal	None			
	volume		volume				
	doi	10.1007/s10579-018-9431-1	doi				
	urls	• http://link.springer.com/content/pdf/10.1007/s10579-018-9431-1.pdf • http://link.springer.com/article/10.1007/s10579-018-9431-1/fulltext.html • http://link.springer.com/content/pdf/10.1007/s10579-018-9431-1.pdf • http://dx.doi.org/10.1007/s10579-018-9431-1	urls	• http://arxiv.org/pdf/1808.09397v1 • http://arxiv.org/abs/1808.09397v1 • http://arxiv.org/pdf/1808.09397v1			
	id	id-743146977200867784	id	id-7452776468122172646			
	abstract		abstract	The wide adoption of electronic health records (EHRs) has enabled a wide range of applications leveraging EHR data. However, the meaningful use of EHR data largely depends on our ability to efficiently extract and consolidate information embedded in clinical text where natural language processing (NLP) techniques are essential. Semantic textual similarity (STS) that measures the semantic similarity between text snippets plays a significant role in many NLP applications. In the general NLP domain, STS shared tasks have made available a huge collection of text snippet pairs with manual annotations in various domains. In the clinical domain, STS can enable us to detect and eliminate redundant information that may lead to a reduction in cognitive burden and an improvement in the clinical decision-making process. This paper elaborates our efforts to assemble a resource for STS in the medical domain, MedSTS. It consists of a total of 174,629 sentence pairs gathered from a clinical corpus at Mayo Clinic. A subset of MedSTS (MedSTS_ann) containing 1,068 sentence pairs was annotated by two medical experts with semantic similarity scores of 0-5 (low to high similarity). We further analyzed the medical concepts in the MedSTS corpus, and tested four STS systems on the MedSTS_ann corpus. In the future, we will organize a shared task by releasing the MedSTS_ann corpus to motivate the community to tackle the real world clinical problems.			
	versions		versions				