cases	doc_1		doc_2		decision	id
			authors	• Goran GlavaÅ¡ • Marc Franco-Salvador • Simone Paolo Ponzetto • Paolo Rosso	DUPLICATES	358
	authors	• GlavaÅ¡, G. • Franco-Salvador, M. • Ponzetto, S. • Rosso, P.	**title**	A Resource-Light Method for Cross-Lingual Semantic Textual Similarity		
	title	A resource-light method for cross-lingual semantic textual similarity	**publication_date**	2018-01-19 15:00:33+00:00		
	publication_date	2018-01-01 00:00:00	**source**	SupportedSources.ARXIV		
	source	SupportedSources.CROSSREF	**journal**	None		
	journal		**volume**			
	volume		**doi**			
	doi	10.1016/j.knosys.2017.11.041	**urls**	• http://arxiv.org/pdf/1801.06436v1 • http://arxiv.org/abs/1801.06436v1 • http://arxiv.org/pdf/1801.06436v1		
	urls	• https://api.elsevier.com/content/article/PII:S0950705117305725?httpAccept=text/xml • https://api.elsevier.com/content/article/PII:S0950705117305725?httpAccept=text/plain • http://dx.doi.org/10.1016/j.knosys.2017.11.041	**id**	id1864959165139727044		
	id	id8675661493242082883	**abstract**	Recognizing semantically similar sentences or paragraphs across languages is beneficial for many tasks, ranging from cross-lingual information retrieval and plagiarism detection to machine translation. Recently proposed methods for predicting cross-lingual semantic similarity of short texts, however, make use of tools and resources (e.g., machine translation systems, syntactic parsers or named entity recognition) that for many languages (or language pairs) do not exist. In contrast, we propose an unsupervised and a very resource-light approach for measuring semantic similarity between texts in different languages. To operate in the bilingual (or multilingual) space, we project continuous word vectors (i.e., word embeddings) from one language to the vector space of the other language via the linear translation model. We then align words according to the similarity of their vectors in the bilingual embedding space and investigate different unsupervised measures of semantic similarity exploiting bilingual embeddings and word alignments. Requiring only a limited-size set of word translation pairs between the languages, the proposed approach is applicable to virtually any pair of languages for which there exists a sufficiently large corpus, required to learn monolingual word embeddings. Experimental results on three different datasets for measuring semantic textual similarity show that our simple resource-light approach reaches performance close to that of supervised and resource intensive methods, displaying stability across different language pairs. Furthermore, we evaluate the proposed method on two extrinsic tasks, namely extraction of parallel sentences from comparable corpora and cross lingual plagiarism detection, and show that it yields performance comparable to those of complex resource-intensive state-of-the-art models for the respective tasks.		
	abstract		**versions**			
	versions					