		doc_1		doc_2	decision	id
cases	authors	Y. Liang - Rui Cao - Jie Zheng - Jie Ren - Ling Gao	authors	- Ling Gao - Jie Ren - Jie Zheng - Rui Cao - Yuxin Liang	DUPLICATES	219
	title	Learning to Remove: Towards Isotropic Pre-trained BERT Embedding	title	Learning to Remove: Towards Isotropic Pre-trained BERT Embedding		
	publication_date	2021-04-12 00:00:00	publication_date	2021-04-12 00:00:00		
	source	SupportedSources.SEMANTIC_SCHOLAR	source	SupportedSources.PAPERS_WITH_CODE		
	journal		journal			
	volume		volume			
	doi	10.1007/978-3-030-86383-8_36	doi			
	urls	https://www.semanticscholar.org/paper/ab151c1ca0479b677003ef200018b93e983aa0ec	urls	- https://arxiv.org/pdf/2104.05274v2.pdf - https://github.com/liangyuxin42/weighted-removal		
	id	id-9088774381427119780	id	id2646592050147452497		
	abstract	None	abstract	Pre-trained language models such as BERT have become a more common choice of natural language processing (NLP) tasks. Research in word representation shows that isotropic embeddings can significantly improve performance on downstream tasks. However, we measure and analyze the geometry of pre-trained BERT embedding and find that it is far from isotropic. We find that the word vectors are not centered around the origin, and the average cosine similarity between two random words is much higher than zero, which indicates that the word vectors are distributed in a narrow cone and deteriorate the representation capacity of word embedding. We propose a simple, and yet effective method to fix this problem: remove several dominant directions of BERT embedding with a set of learnable weights. We train the weights on word similarity tasks and show that processed embedding is more isotropic. Our method is evaluated on three standardized tasks: word similarity, word analogy, and semantic textual similarity. In all tasks, the word embedding processed by our method consistently outperforms the original embedding (with average improvement of 13% on word analogy and 16% on semantic textual similarity) and two baseline methods. Our method is also proven to be more robust to changes of hyperparameter.		
	versions		versions			