

cases	doc_1		doc_2				decision	id
							NOT DUPLICATES	382
	authors	<ul style="list-style-type: none">Haochen TanWei ShaoHan WuKe YangLinqi Song	authors	<ul style="list-style-type: none">Haochen TanWei ShaoHan WuKe YangLinqi Song				
	title	A Sentence is Worth 128 Pseudo Tokens: A Semantic-Aware Contrastive Learning Framework for Sentence Embeddings	title	A Sentence is Worth 128 Pseudo Tokens: A Semantic-Aware Contrastive Learning Framework for Sentence Embeddings				
	publication_date	2022-03-11 00:00:00	publication_date	2022-03-11 00:00:00				
	source	SupportedSources.OPENALEX	source	SupportedSources.INTERNET_ARCHIVE				
	journal	arXiv (Cornell University)	journal					
	volume		volume					
	doi	10.48550/arxiv.2203.05877	doi					
	urls	<ul style="list-style-type: none">https://openalex.org/W4221141171https://doi.org/10.48550/arxiv.2203.05877http://arxiv.org/pdf/2203.05877	urls	<ul style="list-style-type: none">https://web.archive.org/web/20220315003228/https://arxiv.org/pdf/2203.05877v1.pdf				
	id	id-3445149837037963572	id	id-6081150094875659401				
	abstract		abstract	Contrastive learning has shown great potential in unsupervised sentence embedding tasks, e.g., SimCSE. However, We find that these existing solutions are heavily affected by superficial features like the length of sentences or syntactic structures. In this paper, we propose a semantics-aware contrastive learning framework for sentence embeddings, termed Pseudo-Token BERT (PT-BERT), which is able to exploit the pseudo-token space (i.e., latent semantic space) representation of a sentence while eliminating the impact of superficial features such as sentence length and syntax. Specifically, we introduce an additional pseudo token embedding layer independent of the BERT encoder to map each sentence into a sequence of pseudo tokens in a fixed length. Leveraging these pseudo sequences, we are able to construct same-length positive and negative pairs based on the attention mechanism to perform contrastive learning. In addition, we utilize both the gradient-updating and momentum-updating encoders to encode instances while dynamically maintaining an additional queue to store the representation of sentence embeddings, enhancing the encoder's learning performance for negative examples. Experiments show that our model outperforms the state-of-the-art baselines on six standard semantic textual similarity (STS) tasks. Furthermore, experiments on alignments and uniformity losses, as well as hard examples with different sentence lengths and syntax, consistently verify the effectiveness of our method.				
	versions		versions					