cases	doc_1	doc_2	decision	id

doc_1

authors	• Zihao Chen • Hisashi Handa • Kimiaki Shirahama
title	JCSE: Contrastive Learning of Japanese Sentence Embeddings and Its Applications
publication_date	2023-01-19 00:00:00
source	SupportedSources.INTERNET_ARCHIVE
journal	
volume	
doi	
urls	• https://web.archive.org/web/20230120042251/https://arxiv.org/pdf/2301.08193v1.pdf
id	id-3703905663732883978
abstract	Contrastive learning is widely used for sentence representation learning. Despite this prevalence, most studies have focused exclusively on English and few concern domain adaptation for domain-specific downstream tasks, especially for low-resource languages like Japanese, which are characterized by insufficient target domain data and the lack of a proper training strategy. To overcome this, we propose a novel Japanese sentence representation framework, JCSE (derived from "Contrastive learning of Sentence Embeddings for Japanese"), that creates training data by generating sentences and synthesizing them with sentences available in a target domain. Specifically, a pre-trained data generator is finetuned to a target domain using our collected corpus. It is then used to generate contradictory sentence pairs that are used in contrastive learning for adapting a Japanese language model to a specific task in the target domain. Another problem of Japanese sentence representation learning is the difficulty of evaluating existing embedding methods due to the lack of benchmark datasets. Thus, we establish a comprehensive Japanese Semantic Textual Similarity (STS) benchmark on which various embedding models are evaluated. Based on this benchmark result, multiple embedding methods are chosen and compared with JCSE on two domain-specific tasks, STS in a clinical domain and information retrieval in an educational domain. The results show that JCSE achieves significant performance improvement surpassing direct transfer and other training strategies. This empirically demonstrates JCSE's effectiveness and practicability for downstream tasks of a low-resource language.
versions	

doc_2

authors	• Zihao Chen • Hisashi Handa • Kimiaki Shirahama
title	JCSE: Contrastive Learning of Japanese Sentence Embeddings and Its Applications
publication_date	2023-01-19 17:41:46+00:00
source	SupportedSources.ARXIV
journal	None
volume	
doi	
urls	• http://arxiv.org/pdf/2301.08193v1 • http://arxiv.org/abs/2301.08193v1 • http://arxiv.org/pdf/2301.08193v1
id	id-4658880451625687695
abstract	Contrastive learning is widely used for sentence representation learning. Despite this prevalence, most studies have focused exclusively on English and few concern domain adaptation for domain-specific downstream tasks, especially for low-resource languages like Japanese, which are characterized by insufficient target domain data and the lack of a proper training strategy. To overcome this, we propose a novel Japanese sentence representation framework, JCSE (derived from ``Contrastive learning of Sentence Embeddings for Japanese"), that creates training data by generating sentences and synthesizing them with sentences available in a target domain. Specifically, a pre-trained data generator is finetuned to a target domain using our collected corpus. It is then used to generate contradictory sentence pairs that are used in contrastive learning for adapting a Japanese language model to a specific task in the target domain. Another problem of Japanese sentence representation learning is the difficulty of evaluating existing embedding methods due to the lack of benchmark datasets. Thus, we establish a comprehensive Japanese Semantic Textual Similarity (STS) benchmark on which various embedding models are evaluated. Based on this benchmark result, multiple embedding methods are chosen and compared with JCSE on two domain-specific tasks, STS in a clinical domain and information retrieval in an educational domain. The results show that JCSE achieves significant performance improvement surpassing direct transfer and other training strategies. This empirically demonstrates JCSE's effectiveness and practicability for downstream tasks of a low-resource language.
versions	

decision	id
DUPLICATES	8