Cross-lingual Sentence Embedding for Low-resource Chinese-Vietnamese Based on Contrastive Learning

YUXIN HUANG, YIN LIANG, ZHAOYUAN WU, ENCHANG ZHU, and ZHENGTAO YU, Faculty of Information Engineering and Automation, Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, China

Cross-lingual sentence embedding's goal is mapping sentences with similar semantics but in different languages close together and dissimilar sentences farther apart in the representation space. It is the basis of many downstream tasks such as cross-lingual document matching and cross-lingual summary extraction. At present, the works of cross-lingual sentence embedding tasks mainly focus on languages with large-scale corpus. But low-resource languages such as Chinese-Vietnamese are short of sentence-level parallel corpora and clear cross-lingual monitoring signals, and these works on low-resource languages have poor performances. Therefore, we propose a cross-lingual sentence embedding method based on contrastive learning and effectively fine-tune powerful pretraining mode by constructing sentence-level positive and negative samples to avoid the catastrophic forgetting problem of the traditional fine-tuning pre-trained model based only on small-scale aligned positive samples. First, we construct positive and negative examples by taking parallel Chinese Vietnamese sentences as positive examples and non-parallel sentences as negative examples. Second, we construct a siamese network to get contrastive loss by inputting positive and negative samples and fine-tuning our model. The experimental results show that our method can effectively improve the semantic alignment accuracy of cross-lingual sentence embedding in Chinese and Vietnamese contexts.

CCS Concepts: • **Computing methodologies → Natural language processing;**

Additional Key Words and Phrases: Chinese-Vietnamese, low-resource language, cross-lingual sentence embedding, siamese network, mBERT

ACM Reference format:
Yuxin Huang, Yin Liang, Zhaoyuan Wu, Enchang Zhu, and Zhengtao Yu. 2023. Cross-lingual Sentence Embedding for Low-resource Chinese-Vietnamese Based on Contrastive Learning. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 22, 6, Article 176 (June 2023), 18 pages.
https://doi.org/10.1145/3589341

This work was supported by the National Natural Science Foundation of China (grant nos. U21B2027, 61972186, 62266028, 62266027); Yunnan Provincial Major Science and Technology Special Plan Projects (grant nos. 202103AA080015, 202202AD080003); General Projects of Basic Research in Yunnan Province (grant nos. 202201AS070179, 202201AT070915]; Kunming University of Science and Technology "double first-class" joint project (grant no. 202201BE070001-021).
Authors' address: Y. Huang, Y. Liang, Z. Wu, E. Zhu, and Z. Yu (corresponding author), Faculty of Information Engineering and Automation, Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, 727 Jingming South Road, Kunming, Yunnan, China, 650500; emails: huangyuxin2004@163.com, {2712653816, 978105863}@qq.com, ztyu@hotmail.com.

2375-4699/2023/06-ART176 $15.00
https://doi.org/10.1145/3589341

1 INTRODUCTION

The purpose of the cross-lingual sentence embedding task is to encode the sentence semantic information of different languages and map it to a language-independent shared embedding space for alignment, so sentences with similar semantics in different languages have similar vector representation [1], to realize the transmission of semantic information between different languages. As the mainstream method to obtain cross-lingual sentence embedding, the multilingual pre-training model can well capture the syntactic and semantic features in different language sequences [2]. Therefore, it is often used to solve some more complex cross-lingual tasks, such as cross-lingual document matching [3], cross-lingual summary extraction [4].

However, most of the existing sentence embedding work focuses on languages with rich corpora and a large number of co-occurring words (such as English, French [5]), which can help to fine-tune the model. But in languages with scarce parallel sentences corpus such as Chinese-Vietnamese, due to the large syntax differences between languages and the lack of co-occurring words, the pre-training model lacks sufficient anchors, resulting in poor results of previous work [6]. As shown in Figure 1, the semantic correspondence of words in Chinese-Vietnamese sentences $A(S, T)$ does not follow word order correspondence, which leads to a large semantic alignment error in the contextual cross-lingual sentence embeddings learned by the encoder only trained in Chinese and Vietnamese monolingual corpus. And due to the catastrophic forgetting problem, the pre-training model can only focus on one task. It means that when learning the current task B, the knowledge of the previous task A will be lost suddenly.

As shown in Figure 2, Libovický et al. [7] experimentally found that in the mBERT model [8] and XLM, the learned multilingual embeddings are fused in the same semantic space in a seemingly chaotic state, but the embedding spaces corresponding to different languages are offset to varying degrees according to the similarity difference between languages. In language pairs with great differences such as Chinese and Vietnamese, the offset between embedding distributions will affect the semantic similarity calculation of contextual cross-lingual sentence embeddings and reduce the accuracy of the multilingual pre-training model in Chinese Vietnamese sentence semantic alignment task [9].

To solve the above problems, this article proposes a method for fine-tuning Chinese-Vietnamese contextual cross-lingual sentence embeddings by constructing a positive and negative sample and fusing siamese network to alleviate the problem of poor semantic alignment of Chinese-Vietnamese contextual cross-lingual sentence embeddings in multilingual pre-training models due to the scarcity of Chinese-Vietnamese sentence-level parallel corpus and high linguistic variability. We conducted experiments based on mBERT and XLM models, respectively, and the experimental results show that our method can effectively improve the semantic alignment accuracy of cross-lingual sentence embedding in Chinese and Vietnamese contexts.

To summarize, our contributions are as follows:

- Based on the mBERT model and combined with contrastive learning, a cross-lingual sentence embedding fine-tuning model mBERT-SF for Chinese-Vietnamese low-resource bilingual tasks is designed and implemented.
- The effectiveness of the method in improving the semantic alignment of Chinese-Vietnamese contextual cross-lingual sentence embeddings was verified in conjunction with a Chinese-Vietnamese cross-sentence semantic matching task.

2 RELATED WORK

2.1 Multilingual Pre-training Model

Each contextual embedding in a multilingual pre-training model is related to the whole sequence of its input by using the deep learning framework, which can better capture syntactic and

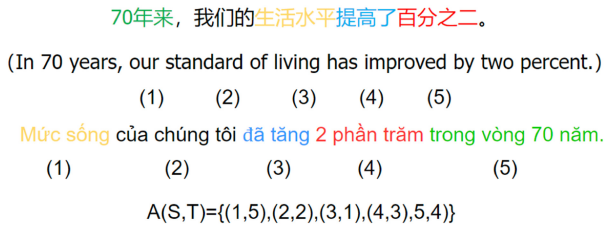

Fig. 1. Differences in grammar and word building between Chinese and Vietnamese.

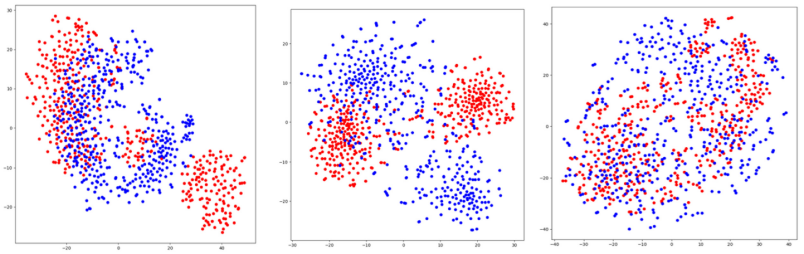

Fig. 2. Deviation between embedding subspaces of Chinese and Vietnamese in mBERT model.

semantic features in input sentences of different languages. Furthermore, the multilingual pre-training model is the mainstream approach to obtaining cross-lingual sentence-level embeddings today.

The multilingual pre-training approach originated from Artetxe et al. [10], who proposed an encoder-decoder framework consisting of a bidirectional LSTM, where the model is pre-trained using a shared BPE word list of 93 languages and a parallel corpus (containing both European and Tanzir languages) to obtain multilingual contextual embeddings. Later, the Rosita model proposed by Mulcaire et al. [11] demonstrated the benefits of multilingual pre-training models on low-resource languages. Papadimitriou et al. [8] proposed **multi-lingual BERT (mBERT)** based on the BERT model [12], which enables the BERT model to be better used for solving cross-lingual tasks by pre-training BERT using Wikipedia data with more than 100 languages. Subsequently, Lample et al. [13] proposed the XLM model to learn multilingual language model by using three pre-training approaches: **Causal language modeling (CLM), Masked language modeling (MLM)**, and **Translation language modeling (TLM)**, which refreshed the best results on many tasks. The XLM-R model [14] further extends the corpus from Wikipedia data to web-wide data on top of XLM, allowing the model to be more rich-resource and validating the significant performance improvement of the large-scale multilingual pre-trained model in cross-lingual migration tasks. And in Liu et al.'s work [15], a plug-and-play embedding generator is introduced to produce the representation of any input token, according to pre-trained embeddings of its morphologically similar ones. Thus, embeddings of mismatch tokens in downstream tasks can also be efficiently initialized. And they get more efficient and better performed downstream NLG models.

2.2 Contrastive Learning

The concept of contrastive learning originates from unsupervised learning. Compared with a supervised learning algorithm, unsupervised learning has no label guidance, and it is more difficult to learn the characteristics of samples in the training process [16, 17]. Contrastive learning is a

data enhancement method. It is mainly to build positive and negative cases and use the contrast loss training model proposed by Hadshell et al. (dimensional reduction by learning an invariant map) to make the distance between positive cases as close as possible and the distance between negative cases as far as possible after mapping to the common semantic space. This method can only use a small amount of data to get better experimental results.

In machine translation, many scholars have applied comparative learning to machine translation. For example, in Gao's work [18], different sentence representation vectors are obtained by masking different words of the same sentence, so only dropout is used as data enhancement to predict the sentence itself for comparative learning. And Yung et al. [19] try to replace the masked part and predict the replaced part by combining the representation vector of the original sentence to obtain an encoder that can fully express the view of the sentence. To make the translated sentences show the characteristics of the author, the article "Towards User-Driven Neural Machine Translation" [20] also introduces the user behavior characteristics in the cache to make the translated text conform to the user's local language habits. They provide a cache-based module, and a user-driven contrastive learning method is proposed to offer NMT the ability to capture potential user traits from their historical inputs under a zero-shot learning fashion based on contrastive learning. And the experimental results confirm that the proposed user-driven NMT can generate user-specific translations.

Based on contrastive learning, Bromley et al. [21] proposed a twin neural network (SNN) for verifying signatures on handwritten input boards, which is a coupling network constructed by two artificial neural networks with the same architecture.

The siamese neural network proposes to input two data into two sub-neural networks as a group and output the corresponding high-dimensional embedding as a representation. It compares the similarity between the two data by calculating the Euclidean distance between the two representation vectors. Therefore, it is often used in image recognition and matching tasks. To better match the input data, the parameters are usually shared between two sub-neural networks. The model is trained by comparing the similarity of two sentence representation vectors calculated by two neural computing networks with the same parameters. This method can effectively reduce the time complexity and greatly shorten the calculation time. It has been proved that this method has an excellent performance in classification tasks, regression tasks, and tasks in the form of Triplet.

Twin neural networks have excellent performance in natural language processing tasks. Nils et al. [5] used twin networks to encode English sentences and fine-tune BERT and RoBERTa models in combination with various tasks. Wang et al. [22] also used twin networks to calculate the similarity of Chinese sentences.

3 CONTEXT-BASED CHINESE-VIETNAMESE CROSS-LINGUAL SENTENCE EMBEDDING MODEL

To sum up, based on the multilingual pre-training model mBERT, this article proposes a Chinese-Vietnamese contextual cross-lingual sentence embedding fine-tuning method integrating the siamese network (mBERT-SF). By building a fine-tuning layer with the idea of a siamese network in the image processing field, the contextual cross-lingual sentence embeddings obtained in the mBERT model are reconstructed to achieve higher semantic similarity, better alignment, and more reasonable distribution among the Chinese-Vietnamese sentence embeddings, thus alleviating the semantic alignment bias problem of the mBERT model on language pairs with large linguistic differences like Chinese-Vietnamese. The overall architecture of the model is shown in Figure 3.

The context-based Chinese-Vietnamese cross-lingual sentence embedding model mainly consists of a Chinese-Vietnamese contextual cross-lingual sentence embedding acquisition layer based on the mBERT model and a fine-tuning layer constructed by a siamese network. The model

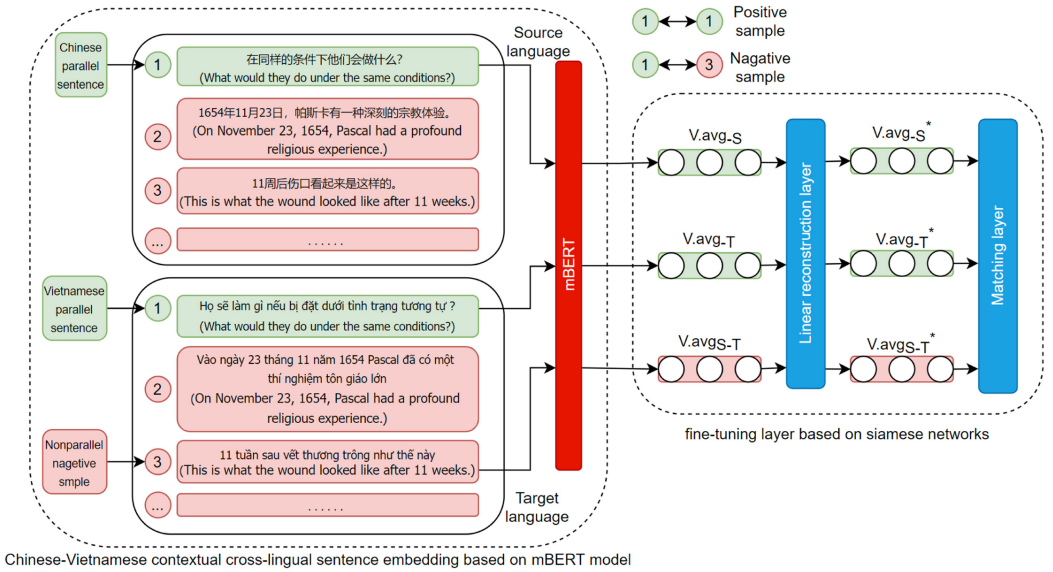

Fig. 3. Context-based Chinese-Vietnamese cross-lingual sentence embedding model.

first inputs the constructed Chinese-Vietnamese parallel sentence pairs and randomly selected Vietnamese sentences into the mBERT model to obtain the corresponding contextual sentence embeddings, where $V.avg_{-S}$ represents the average value of the vectors corresponding to the Chinese word segmentation result in mBERT, $V.avg_{-T}$ represents the average value of the vectors corresponding to the Vietnamese word segmentation result in mBERT in the positive sample composed of parallel sentence pairs, and $V.avg_{S-T}$ represents the average value of the vectors corresponding to the Vietnamese word segmentation result in the contextual embedding of randomly selected Vietnamese sentences in the negative sample. After the linear reconstruction layers fusing the siamese network structure, the fine-tuned contextual cross-lingual sentence embeddings $V.avg_{-S}{}^{*}$, $V.avg_{-T}{}^{*}$, and $V.avg_{T-N}{}^{*}$ are, respectively, obtained. Then, fed them to the matcher for loss calculation to reverse optimize the linear reconstruction layers.

3.1 Chinese-Vietnamese Contextual Cross-lingual Sentence Embedding based on mBERT Model

In the cross-lingual sentence embedding learning process, the mBERT model takes monolingual sentences as input and adds [CLS] and [SEP] tags at the beginning and end of the sentences, respectively, to delimit the sentence extent. Unlike the word-splitting mechanism in traditional static embedding methods [23], the mBERT model uses a sub-word list with language unit granularity between characters and words to segment the input sentences. And Chinese is divided by word as the smallest unit, while Vietnamese is divided by word or syllable as the smallest unit, as shown in Table 1. Where "##" means that the syncopated character or word needs to be connected to the preceding character.

Take the processing of Chinese input sentences as an example: The mBERT model divides "我爱吃苹果" into five individual characters by finding the sub-word word list and generates a corresponding one-dimensional word vector for each character. In addition, a text vector for portraying the global semantic information of the input sequence is automatically learned based on the [CLS] and [SEP] labels added to the model and fused with other word vectors for distinguishing the

Table 1. Example of Chinese Vietnamese Sentence Pair Segmentation in mBERT Model

	Chinese Segment	Vietnamese Segment
Input	我爱吃苹果(I like eating apples)	tôi thích ăn táo
Word list	我##爱##吃##苹##果(I ##like ##eating ##apples)	tôi thích ăn tá ##o
Output	我爱吃苹果(I like eating apples)	tôi thích ăn tá ##o
	……	……

content information represented by different sentences. Finally, because of the difference in semantic information carried by characters or words appearing at different positions in the text (e.g., "我爱你" and "你爱我"), the mBERT model also appends a position vector to the word vectors at different positions to differentiate them. In the final output layer, the model generates a corresponding 768-dimensional vector as a representation for each of the sliced word vectors and labels.

For the acquisition of Chinese-Vietnamese contextual cross-lingual sentence embedding, this article tests two methods, respectively: One method is to directly use the output vector corresponding to the [CLS] layer in the mBERT output as the representation of the entire sentence. The [CLS] tag can further capture contextual information in the input sequence through the Self-Attention mechanism [24] and is therefore commonly used as a sentence-level semantic representation. The other is the average vector method, that is, the average vector of the word vector in the seventh hidden layer of the mBERT model is used as the representation of the input sentence. The average vector method can capture the representation of each word in the sentence, so it is also a common method for sentence representation [5].

3.2 Linear Fine-tuning Layer based on Siamese Networks

Although the Chinese-Vietnamese contextual cross-lingual sentence embedding obtained based on the mBERT model can extract semantic information from each word in the text, the lack of parallel corpus as a supervisory signal in the model training process leads to a certain alignment bias in the distribution of the learned contextual sentence embeddings with language variability. Therefore, a cross-lingual sentence embedding fine-tuning layer integrating the siamese network structure is proposed to reconstruct the contextual cross-lingual sentence embedding of Chinese and Vietnamese to maximize the similarity between the same semantic embedding and improve the semantic alignment accuracy of the mBERT model in Chinese-Vietnamese cross-lingual tasks. The overall structure of the fine-tuning layer is shown in Figure 4.

The fine-tuning layer consists of two subnetworks Network1/Network2 with the same structure and a matcher. Based on the idea of the linear mapping in static cross-lingual word embedding, the linear reconstruction layer composed of two subnetworks reconstructs the Chinese sentence embedding $V.avg\_S$ and Vietnamese sentence embedding $V.avg\_T$ corresponding to the Chinese and Vietnamese input sentence pairs, respectively. Each subnetwork is composed of a full connection layer and a Dropout layer [25]. Among them, the size of the full connection layer is 768 * 768 dimensions, which is responsible for the feature extraction of the original context sentence embedding output by the mBERT model. To further improve the generalization ability of the model, a Dropout layer is added after the full connection layer FC to prevent the overfitting problem of the model by randomly eliminating the neurons in the full connection layer with a probability p.

The two sub-networks Network1 and Network2 feature extraction process is shown in Equation (1). Since the structure of the two networks is the same and the weights are shared, we use the same formula here to show the computational process of both. x represents the contextual cross-lingual sentence embedding $V.avg\_S$, $V.avg\_T$ before the fine-tuning of Chinese or

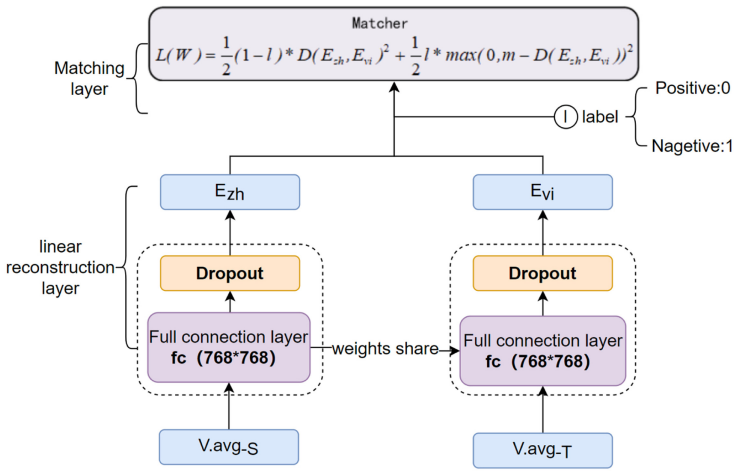

Fig. 4. Linear fine-tuning layer integrated with siamese network structure.

Vietnamese.

$$y = pf(Wx) \tag{1}$$

In Equation (1), y represents the output is reconstructed by subnetworks Network1 and Network2, where $pf(Wx)$ represents the output of the dropout layer, p is the random rejection probability of neurons, and W is the weight value of fully connected layer FC. The final result y can represent the reconstructed Chinese and Vietnamese contextual sentence embeddings E_{zh} and E_{vi}.

Since the training goal of the subnetworks in the model is to make the contextual sentence embeddings in the positive samples as similar as possible and the sentence embeddings between the negative samples as not similar as possible after feature extraction, this article uses contrastive loss as the fine-tuning criterion to fine-tune the two sub-networks in reverse, as shown in Equation (2).

$$L(W) = \frac{1}{2}(1-l) * D(E_{zh}, E_{vi})^2 + \frac{1}{2} l * \max(0, m - D(E_{zh}, E_{vi}))^2, \tag{2}$$

$$D(E_{zh}, E_{vi}) = \|E_{zh} - E_{vi}\|_2, \tag{3}$$

where E_{zh} and E_{vi} are Chinese and Vietnamese contextual cross-lingual sentence embeddings reconstructed by the fine-tuning layer. $D(E_{zh}, E_{vi})$ represents the Euclidean distance between the two embeddings, as shown in Equation (3). l represents the label corresponding to the input Chinese and Vietnamese sentence pair when the input is a positive sample constructed from parallel sentence pairs, $l = 0$ when it is a negative sample composed of non-parallel sentence pairs, $l = 1$. m is the maximum margin value of margin, and by operating $m - D(E_{zh}, E_{vi})$, a smaller loss can be generated for the pairs of sentences in the negative samples where the Euclidean distance exceeds the margin value to satisfy the optimization objective of the model.

4 EXPERIMENT

4.1 Dataset

Since English belongs to the Indo-European language family along with many European languages and is the largest language system in the world, the current training set for fine-tuning multilingual pre-training models is mostly the English dataset. However, the grammatical structure and word formation rules of Chinese and Vietnamese are different from those of English, so English is not

Table 2. Examples of Chinese-Vietnamese Parallel Sentence Pairs

Chinese Sentence	Vietnamese Sentence	label
在同样的条件下他们会做什么?	Họ sẽ làm gì n´u bị đt d´i tình trạng tng t¸?	0
怎么会不一样呢?	Làm th´ nào nó có th khác đi ?	0
……	……	…

Table 3. Training Set Examples after Adding Chinese-Vietnamese Non-parallel Sentence Pairs

Chinese Sentence	Vietnamese Sentence	label
在同样的条件下他们会做什么?	Họ sẽ làm gì n´u bị đt d´i tình trạng tng t¸?	0
在同样的条件下他们会做什么?	Tôi sẽ ki´m đc bao nhiêu n´u gian lận ?	1
怎么会不一样呢?	Làm th´ nào nó có th khác đi ?	0
怎么会不一样呢?	Căn ban đó là 1 v´n đ` v` dc phm tái tạo ?	1
……	……	…

the best choice for fine-tuning the Chinese-Vietnamese cross-lingual embedding. And the model needs to construct a small-scale Chinese-Vietnamese parallel sentence pair dataset as the training set for the fine-tuning method in this article first.

The Chinese-Vietnamese parallel sentence pairs in the dataset are mainly from Wikipedia, which is the current mainstream encyclopedia website with entries in over 100 languages around the world. Although there is still a large gap between the Vietnamese corpus and resource-rich languages such as Chinese, there are still many paragraph-level comparable corpora, most of which are derived from encyclopedic entries on the same topic, and their semantic contents are very close. Using web crawler technology, about 20,000 sizes of Chinese-Vietnamese pseudo-parallel sentence pairs were extracted from Wikipedia. To ensure the quality of the sentences, Chinese and Vietnamese sentences with some words less than 5 were removed, respectively, then the wrong sentence pairs containing special characters were eliminated using regularization technology, and finally, the 2016 pairs with the highest semantic similarity were filtered out from the remaining 7,000 pairs as positive samples. In addition, 448 Chinese-Vietnamese parallel sentence pairs were also manually refined and labeled as a test set to ensure the validity of the model effect evaluation, and the data of parallel sentence pairs were constructed in the format shown in Table 2, with a label of 0 representing the data as parallel sentence pairs.

Due to the limited scale of positive sample data construction, to further improve the model's ability to discriminate different semantic sentence embeddings, negative sample data of the same scale as positive samples are also constructed as the training set. The 2016 negative sample sentence pairs were constructed by randomly selecting Vietnamese-translated sentences for the Chinese sentences in the positive sample that also belonged to the positive sample but did not correspond semantically, and mixing them with the positive sample, the format of the constructed data is shown in Table 3, and the data label used for the negative sample is 1.

The size of each dataset is shown in Table 4.

4.2 Experimental Parameter Setting

The parameters of the fine-tuning layer fusing the siamese network are set as follows: The Chinese-Vietnamese contextual cross-lingual sentence embedding dimension as input is 768 dimensions, and the output embedding dimension after fine-tuning is constant. The random

Table 4. The Scale of Datasets

Dataset	Scale
Training Set	4,032
Test Set	448

masking probability of the Dropout layer neurons p is 0.2 for the two sub-networks in the siamese network. The maximum margin value m in the matcher consisting of contrastive loss is set to 1.0. Model optimization using the Adam optimizer with a learning rate lr of 5e-6, a batch_size of 64 for the training samples, and epochs of 70 for the iterative rounds. The normalization parameters performed for the Chinese-Vietnamese contextual cross-lingual sentence embedding after fine-tuning the output are ["unit", "center", "unit"], where "unit" stands for length normalization and "center" stands for centering operation. The exploration experiments about the optimal parameter settings of the model are detailed in Section 3.5.

4.3 Evaluating Indicator

In the Chinese-Vietnamese cross-lingual sentence semantic matching task, the model uses cosine similarity to measure the semantic similarity between two contextual cross-lingual sentence embeddings. The model uses the accuracy of sentence alignment P@N (the semantic alignment accuracy when N Vietnamese sentences are selected as candidates) as a measure of the model's semantic alignment effectiveness, which is calculated as shown in Equation (4).

$$P = \frac{\sum_i^T \|C(E_{zh})\|}{T} \times 100\%, \tag{4}$$

where T represents the size of the Chinese-Vietnamese sentence pairs in the test set, and $C(E_{zh})$ represents the list of Vietnamese sentence candidates retrieved by the model based on the cosine similarity for the Chinese contextual cross-language sentence embedding E_{zh}, and takes 1 if the set of candidate sentences $C(E_{zh})$ contains the correct Vietnamese translation, otherwise takes 0.

The cosine similarity is calculated as shown in Equation (5), where E_{vi} is the Vietnamese contextual cross-lingual sentence embedding in the test set.

$$\cos(E_{zh}, E_{vi}) = \frac{E_{zh}E_{vi}}{\|E_{zh}\|_2 \times \|E_{vi}\|_2}. \tag{5}$$

4.4 Baseline Model

To highlight the effectiveness of this article's fine-tuning approach in alleviating the problem of contextual cross-lingual sentence embedding alignment bias of mBERT models due to Chinese-Vietnamese language variability, three current mainstream multilingual pre-training models were selected as baselines and tested for comparison on Chinese-Vietnamese cross-lingual sentence semantic matching task and visual embedding distribution experiment, and each baseline model is described as follows:

- mBERT model: the **Multilingual BERT (mBERT)** model proposed by Devlin et al. [8], based on the original BERT model, uses the **masked language modeling (MLM)** method to pre-train on Wikipedia data composed of 104 languages.
- XLM model: a multilingual pre-training model proposed by Lample et al. [13], which introduces two additional pre-training methods, **Causal language modeling (CLM)** and **Translation language modeling (TLM)** based on BERT. The XLM model is also trained on the

Table 5. Effects of Different Pre-training Models on Chinese-Vietnamese Sentence Semantic Matching Task

Model	P@1	P@5
XLM	58.0%	76.7%
mBERT	41.2%	62.5%
mT5	37.8%	59.2%
XLM-R	47.2%	67.8%
XLM-SF	**70.1%**	**85.7%**
mBERT-SF	63.1%	81.7%
mT5-SF	45.1%	65.2%
XLM-R-SF	54.1%	73.7%

Wikipedia corpus covering more than 100 languages, but the XLM model has more parameters and a larger shared word list than the mBERT model.

- mT5 model: a multilingual variant of T5 [26] proposed by Linting et al. [27]. And it was pre-trained on a new Common Crawl-based dataset covering 101 languages. mT5 inherits all of the benefits of T5, such as its general-purpose text-to-text format, its design based on insights from a large-scale empirical study, and its scale.
- XLM-R model: The XLM-RoBERTa Large model proposed by Conneau et al. [28] on the basis of the XLM model is a large-scale multilingual pre-training model that uses the entire network data with a larger amount of data than Wikipedia as training corpus.

4.5 Analysis of Experimental Results

4.5.1 Chinese-Vietnamese Cross-lingual Sentence Semantic Matching Task Review. To verify the effectiveness of this article's fine-tuning method on the multilingual pre-training model mBERT, a set of comparative experiments is designed to compare the Chinese and Vietnamese contextual cross-lingual sentence embeddings directly outputted by mainstream multilingual pre-training models such as mBERT and sentence embeddings outputted after linear fine-tuning on the Chinese and Vietnamese cross-lingual sentence semantic matching task, respectively. The task first normalizes the contextual cross-lingual sentence embeddings corresponding to the Chinese-Vietnamese sentence pairs in the test set and then uses cosine similarity to find semantically corresponding N Vietnamese sentence embeddings as candidates for matching for each Chinese sentence embedding in the test set, and the semantic matching accuracy P@N results of different models at N candidate sentences are shown in Table 5.

By analyzing the experimental results in the table, it can be seen that the Chinese-Vietnamese contextual cross-lingual sentence embeddings obtained by the fine-tuned mBERT-SF method can improve the semantic alignment accuracy of Chinese-Vietnamese contextual cross-lingual sentence embeddings compared with the original model like mBERT, XLM et al. mainstream multilingual pre-training models in the Chinese-Vietnamese cross-lingual sentence semantic matching tasks @1 and @5, which fully demonstrates that the mBERT-SF method can effectively improve the semantic alignment accuracy of Chinese-Vietnamese contextual cross-lingual sentence embeddings using only small-scale Chinese-Vietnamese parallel sentence pairs and alleviate the semantic bias problem caused by the scarcity of Chinese-Vietnamese sentence-level parallel corpus in the mBERT model. Meanwhile, by comparing the alignment accuracy of different models on the Chinese-Vietnamese cross-lingual sentence semantic matching task, it can be seen that the effect of

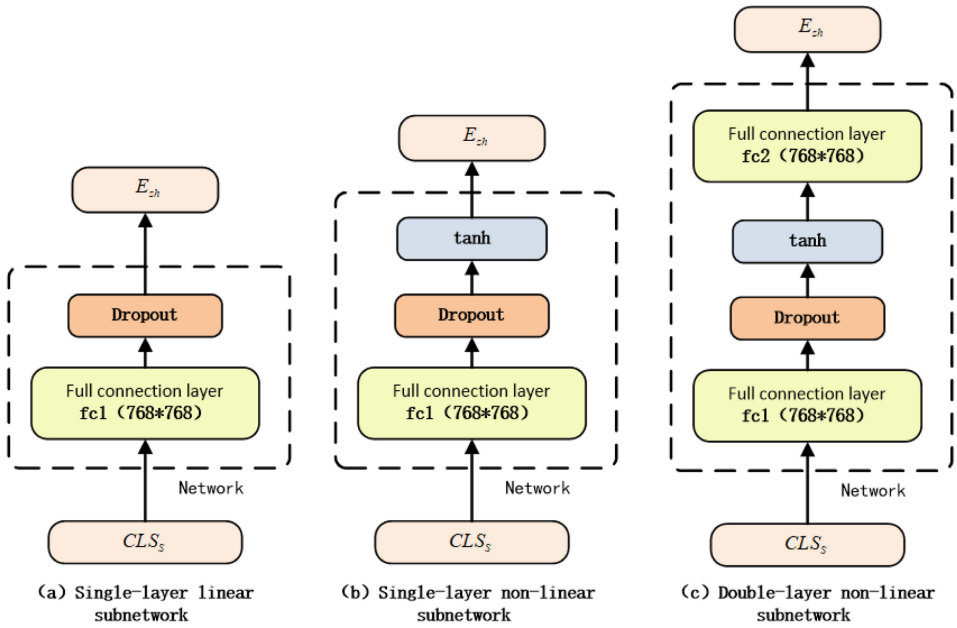

Fig. 5. Alignment accuracy of models based on different sub-network structures.

the XLM model > mBERT model, and the effect decreases gradually with the increase of model data volume and the number of parameters. It is speculated that this is due to the uneven amount of data used in model training and the variability of different languages. Vietnamese, as a low-resource language, has a significantly lower percentage of the training corpus than resource-rich languages such as English, resulting in the cross-lingual knowledge learned by the model being more biased toward Indo-European languages and poorly migrated to languages with larger linguistic differences and smaller scales such as Vietnamese. This also corroborates that the Chinese-Vietnamese contextual cross-lingual sentence embedding reconstructed by the siamese network fine-tuning layer eliminates the influence of language variability on the semantic similarity calculation and preserves the semantic information as better as possible, which makes the mBERT model applicable to cross-lingual tasks on low-resource language pairs with large differences such as Chinese and Vietnamese.

4.5.2 Effect of Different Fine-tuning Layer Structures on Model Effects. In the fourth group of experiments, to explore the optimal fine-tuning layer architecture, three different structures of subnetwork were constructed as reconfiguration layers in the fine-tuning layer, and the three structures are shown in Figure 5.

Among them, the (a) figure is a single-layer linear reconstruction network constructed based on the linear idea, consisting of a fully connected layer fc1 and a Dropout layer; (b) figure is a single-layer nonlinear network structure constructed by adding an activation function layer on the base of the single-layer linear reconstruction network, and after comparing the model effect on Relu function and tanh function, the tanh function is finally selected as the activation function layer optimal setting. Figure 5(c) is a two-layer nonlinear reconstruction network composed of a new fully connected layer fc2 on the base of Figure 5(b). The best experimental results of the three subnetworks structure after 10 rounds of iterative training with the same optimal parameter settings on the test sets of cross-lingual sentence semantic matching tasks @1 and @5 are shown in Table 6.

Table 6. Alignment Accuracy of Models Based on Different Sub-network Structures

Model	P@1	P@5
linear single-layer	**63.1%**	**81.7%**
nonlinear single-layer(tanh)	53.3%	72.3%
nonlinear single-layer(ReLU)	35.5%	57.4%
nonlinear double-layer(tanh)	52.9%	73.8%

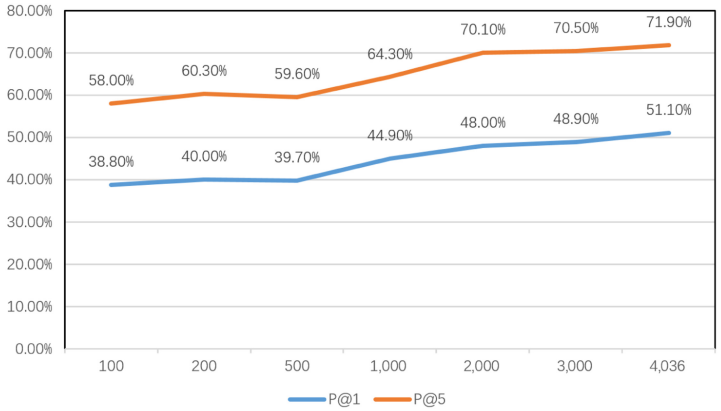

Fig. 6. Alignment accuracy of models based on different data volume.

By analyzing the data in the table, it can be seen that the single-layer linear structure achieves the best alignment effect compared with the nonlinear structure. Therefore, the model finally adopts the linear structure as the architecture of the two subnetworks in the reconstruction layer. In the comparison of nonlinear structures, the alignment accuracy of the model decreases with the increase in model complexity. It is speculated that this is because the mapping process of two full connection layers will lead to the loss of semantic information in the original contextual cross-lingual sentence embedding. While the single-layer network structure can reconstruct the Chinese-Vietnamese contextual cross-lingual sentence embedding, it can also retain the semantic information extracted from the original embedding to a great extent. Therefore, the single-layer linear structure is finally selected as the sub-network architecture of the reconstruction layer in this method.

4.5.3 Effect of Different Data Volume on Model Effects. To explore whether the amount of data has an impact on the effect of the model, the fifth group of experiments was divided into seven groups according to the amount of data used in the training model. Different amounts of data were input into the model with the same parameters. The best experimental results on Chinese Vietnamese cross language sentence meaning matching tasks @1 and @5 in the test set are shown in Figure 6.

By analyzing the data in the table, it can be seen that when the amount of experimental data is less than 1,000, the alignment result is too low and unstable due to too little data volume. When the amount of experimental data is greater than 1,000, the alignment accuracy of the model will increase with the increase of experimental data. The rising rate will slow down after the amount of experimental data reaches 3,000.

Table 7. Alignment Accuracy of Models Based on Different Measurement Methods

measurement methods	P@1	P@5
Euclidean distance	**63.1%**	**81.7**%
Cosine distance	27.5%	40.6%
Manhattan distance	59.9%	74.3%

Table 8. Alignment Accuracy of Models Based on Different Sentence Vector Value Methods

sentence vector value methods	P@1	P@5
CLS	40.17%	58.04%
$V.avg_{-12thlayer}$	54.9%	75.2%
$V.avg_{-11thlayer}$	58.2%	75.0%
$V.avg_{-10thlayer}$	60.3%	78.3%
$V.avg_{-9thlayer}$	60.9%	79.1%
$V.avg_{-8thlayer}$	62.7%	80.6%
$V.avg_{-7thlayer}$	**63.1%**	**81.7%**
$V.avg_{-6thlayer}$	62.7%	81.0%
$V.avg_{-5thlayer}$	57.6%	77.7%
$V.avg_{-4thlayer}$	51.1%	74.8%
$V.avg_{-3rdlayer}$	50.8%	69.9%
$V.avg_{-2ndlayer}$	48.7%	70.5%
$V.avg_{-1stlayer}$	45.8%	67.2%
$V.avg_{-alllayer}$	60.5%	80.1%

4.5.4 Effect of Different Measurement Methods on Model Effects. To explore the optimal measurement method, the sixth group of experiments tested the models using Euclidean distance, cosine distance, and Manhattan distance, respectively. The sentence meaning matching results are shown in Table 7.

By analyzing the data in the table, it can be seen that under the optimal parameter setting, the alignment accuracy obtained by using Euclidean distance measurement is the highest, while the result using Manhattan distance is not as good as Euclidean distance. The use of cosine similarity will lead to a worse effect on the model. This is because the cosine similarity is calculated by calculating the angle between the two sentence vectors to calculate the difference between sentence vectors. This will lead to the cosine distance being more distinguishing from the direction difference and insensitive to the absolute value, so the sentence vectors with similar directions but different distances in the semantic space are misjudged as similar sentence vectors, resulting in poor sentence meaning-matching effect.

4.5.5 Effect of Different Sentence Vector Value Methods on Model Effects. To explore the optimal method of sentence vector selection, we tested the effect of the average value of the CLS tag and the vector at the different layers of the mBERT hidden layer on the task of Chinese Vietnamese cross-lingual semantic matching. The experimental results are shown in Table 8.

It can be seen from the table that the experiment results in the middle layer network are the best. The reason is that the mBERT model encodes rich linguistic-level information: The surface information features are in the bottom layer network, the syntax information features are in the middle layer network, and the semantic information features are in the upper layer network [29].

Table 9. Alignment Accuracy of Models Based on Different Inputs

inputs	P@1	P@5
parallel corpus	55.1%	75.7%
parallel + non-parallel corpus	**63.1%**	**81.7%**

This means that the semantic vectors obtained from the middle layer network of the mBERT model can better represent the syntactic features of sentences in different languages. Compared with the word meaning information displayed by the underlying network and the overall semantic features displayed by the high-level network, the syntactic information features are more effective for sentence matching tasks.

4.5.6 Effect of Training on the Model by Using Only Parallel Sentences. To prove that non-parallel corpora have positive help for model training, the seventh group of experiments compares the training of models using only parallel corpora with the training of parallel corpora and non-parallel corpora at the same time. The results are shown in Table 9.

It can be seen from the results in the table that non-parallel corpora are of positive significance to the task of Chinese Vietnamese cross-lingual sentence meaning matching.

4.5.7 Visualization Comparison before and after Fine-tuning of Chinese-Vietnamese Contextual Cross-lingual Sentence Embedding. To visualize the bias of embedding distribution in the multilingual pre-training models due to Chinese-Vietnamese linguistic differences and the changes in the distribution of Chinese-Vietnamese contextual cross-lingual sentence embeddings before and after fine-tuning, the contextual cross-linguistic sentence embeddings corresponding to the Chinese-Vietnamese parallel sentence pairs in the test set in pre-training models such as mBERT and XLM were downscaled to 2-dimensional embeddings. And using the matplotlib tool to visualize the embedding distribution. Figure 7 shows the embedding distributions of Chinese and Vietnamese contextual cross-language sentences after visualization with different multilingual pre-training models and the mBERT-SF method. Among them, red and yellow represents Chinese sentence embedding and blue represents Vietnamese sentence embedding. We use the average value of the sentence representation embedding obtained from all layers of the model to draw the figure.

Figures 7(a) and (c) show the cross-lingual sentence embedding distributions of the XLM and mBERT models, respectively. It can be seen that the Chinese and Vietnamese sentence embedding distributions in the three models have obvious deviations, and the embedding subspaces of the two do not overlap due to the Chinese and Vietnamese language differences. In Figures 7(b) and (d), after the reconstruction of the siamese network fine-tuning layer, the overlap between Chinese and Vietnamese sentence embedding spaces in the mBERT-SF and XLM-SF model is significantly improved and the distribution is more uniform, which makes it easier to perform the semantic similarity calculation between sentence embeddings and improve the accuracy of Chinese-Vietnamese contextual cross-lingual sentence embedding in cross-lingual semantic alignment.

4.5.8 Influence of Different Languages on Model Effect. To prove that our method works in different languages, we tested the effect of the model on Chinese-English and English-French, respectively. Among them, Chinese and English do not belong to the same etymology; there are no co-occurrence words and great grammatical differences. English and French belong to the same etymology. The two languages have similar grammar and some words have similar meanings. We used 2,000 pairs of parallel sentences and 2,000 pairs of non-parallel sentences in each language as training sets. The experimental results are shown in Table 10.

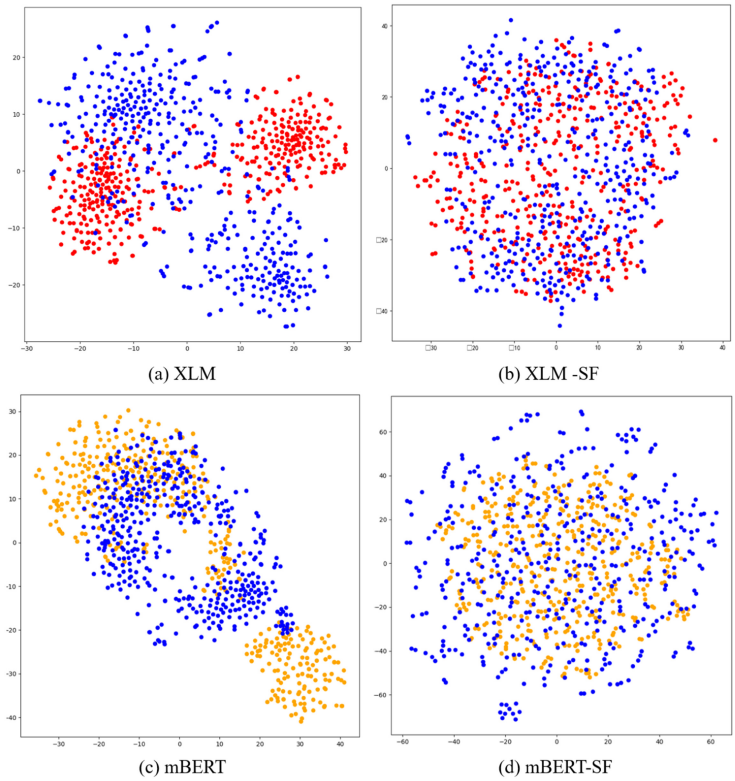

Fig. 7. Chinese-Vietnamese contextual cross-lingual sentence embedding distribution in different models.

Table 10. Alignment Accuracy of Models Based on Different Input Language

Input Language	mBERT		mBERT-SF	
	P@1	P@5	P@1	P@5
Chinese-English	71.1%	85.7%	73.0%	86.3%
English-French	76.8%	88.7%	78.5%	90.1%
Chinese-Vietnamese	41.2%	62.5%	63.1%	81.7%

From the analysis of the results in the table, it can be concluded that Chinese-English and English-French, which contain rich expectations, use only a small amount of training corpus, and the matching effect of both the mBERT and mBERT-SF models is better than that of Chinese-Vietnamese. It can also be seen that the matching effect of mBERT-SF in different languages has improved compared with mBERT. And we also tested the performance of our model on these two pairs of languages with rich training corpus (20,000 pairs of parallel sentences and 20,000 pairs of non-parallel sentences). The experimental results are shown in Table 11.

The experimental results can prove that when our method has a rich corpus as training data, whether the target language belongs to the same language family or not, it has a certain improvement on the baseline model.

4.5.9 Effect on Document-matching Task. To prove that our method is not only effective on sentence-level matching tasks, we test the effect of the model on document-matching tasks. We

Table 11. Alignment Accuracy of Models Based on Different Input Language

Input Language	mBERT		mBERT-SF	
	P@1	P@5	P@1	P@5
Chinese-English	82.3%	87.7%	83.5%	89.1%
English-French	85.7%	90.1%	86.5%	90.9%

Table 12. Effect on Document-matching Task

model	P@1	P@5
mBERT	33.1%	45.7%
mBERT-SF	37.5%	52.3%

Table 13. Effect on Different Ways to Get Negative Examples

negative examples	P@1	P@5
random	58.6%	75.4%
worst	**63.1%**	**81.7%**
average	57.3%	76.5%

crawled 500 parallel paragraph texts from Vietnamese-Chinese news websites, each of which contains 30 to 50 words. Based on these paragraphs, we have constructed 500 non-parallel Chinese-Vietnamese paragraphs in a disordered way. We tested the effect of the document-matching task on this dataset, and the results are shown in Table 12.

From the analysis of the results in the table, we can see that our method also has some improvement compared with the baseline model in the text-matching task. This proves that our method can also achieve good results in downstream tasks.

4.5.10 Effect on Different Ways to Get Negative Examples. There are three strategies in the selection of negative cases in contrastive learning when calculating the loss function: random negative example, worst negative example, and average value. We tested the effect of each method in the model, and the experimental results are shown in Table 13.

The experimental results show that the best experimental results can be obtained when the worst calculation result is used as the negative example. The reason is that the negative example we use is not the perfect negative example. When we construct the negative example, we match the Chinese sentence with the random Vietnamese sentence, which will lead to the situation that the Vietnamese sentence as the negative example is similar to the Chinese sentence. This will lead to the fact that the positive and negative cases cannot be separated far enough in the semantic space so the model cannot distinguish the difference between positive and negative cases. Therefore, using the worst negative example can eliminate the negative impact of these semantically similar negative examples.

5 CONCLUSION

Aiming at the problem of semantic alignment deviation in Chinese-Vietnamese context cross-language sentence embedding due to language similarity difference in multilingual pre-trained model mBERT, a fine-tuning method of Chinese Vietnamese context cross-language sentence embedding based on a twin network is proposed. The experimental results show that the Chinese

Vietnamese context cross-lingual sentence embedding improved by the twin tuning method achieves the best effect in the cross-lingual sentence meaning matching task compared with other baseline models, which fully proves that the tuning method in this article can effectively improve the semantic alignment accuracy of Chinese-Vietnamese context cross-language sentence embedding, alleviating the alignment deviation caused by Chinese Vietnamese language differences in mBERT model. In the next step, we can consider introducing the pretraining method of prompt to realize the accuracy of embedding cross-lingual sentences in semantic alignment in the training stage of the multilingual pretraining model.

REFERENCES

[1] Quoc V. Le and Tomás Mikolov. 2014. Distributed representations of sentences and documents. *CoRR* abs/1405.4053 (2014).

[2] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2017. Unsupervised learning of sentence embeddings using compositional n-gram features. *CoRR* abs/1703.02507 (2017).

[3] Ahmed El-Kishky and Francisco Guzmán. 2020. Massively multilingual document alignment with cross-lingual sentence-mover's distance. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 616–625. Retrieved from https://aclanthology.org/2020.aacl-main.62/.

[4] Yu Bai, Yang Gao, and Heyan Huang. 2021. Cross-lingual abstractive summarization with limited parallel resources. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 6910–6924. DOI: http://dx.doi.org/10.18653/v1/2021.acl-long.538

[5] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 3980–3990. DOI: http://dx.doi.org/10.18653/v1/D19-1410

[6] Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. In *Proceedings of the 7th International Conference on Learning Representations*. OpenReview.net. Retrieved from https://openreview.net/forum?id=SkEYojRqtm.

[7] Jindrich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics (Findings of ACL)*. Association for Computational Linguistics, 1663–1674. DOI: http://dx.doi.org/10.18653/v1/2020.findings-emnlp.150

[8] Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2522–2532. DOI: http://dx.doi.org/10.18653/v1/2021.eacl-main.215

[9] Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Conference of the Association for Computational Linguistics*. Association for Computational Linguistics, 4996–5001. DOI: http://dx.doi.org/10.18653/v1/p19-1493

[10] Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Trans. Assoc. Comput. Ling.* 7 (2019), 597–610. Retrieved from https://transacl.org/ojs/index.php/tacl/article/view/1742.

[11] Phoebe Mulcaire, Jungo Kasai, and Noah A. Smith. 2019. Polyglot contextual representations improve crosslingual transfer. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 3912–3918. DOI: http://dx.doi.org/10.18653/v1/n19-1392

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 4171–4186. DOI: http://dx.doi.org/10.18653/v1/n19-1423

[13] Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR* abs/1901.07291 (2019).

[14] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 8440–8451. DOI: http://dx.doi.org/10.18653/v1/2020.acl-main.747

[15] Xin Liu, Baosong Yang, Dayiheng Liu, Haibo Zhang, Weihua Luo, Min Zhang, Haiying Zhang, and Jinsong Su. 2021. Bridging subword gaps in pretrain-finetune paradigm for natural language generation. *CoRR* abs/2106.06125 (2021).
[16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. *CoRR* abs/2002.05709 (2020).
[17] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *CoRR* abs/2004.11362 (2020).
[18] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. *CoRR* abs/2104.08821 (2021).
[19] Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James R. Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 4207–4218. DOI: http://dx.doi.org/10.18653/v1/2022.naacl-main.311
[20] Huan Lin, Liang Yao, Baosong Yang, Dayiheng Liu, Haibo Zhang, Weihua Luo, Degen Huang, and Jinsong Su. 2021. Towards user-driven neural machine translation. *CoRR* abs/2106.06200 (2021).
[21] Jane Bromley, James W. Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "Siamese" time delay neural network. *Int. J. Pattern Recog. Artif. Intell.* 7, 4 (1993), 669–688. DOI: http://dx.doi.org/10.1142/S0218001493000339
[22] Zhen Wang, Xiangxie Zhang, and Yicong Tan. 2021. Chinese sentences similarity via cross-attention based siamese network. *CoRR* abs/2104.08787 (2021).
[23] Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated p-mean word embeddings as universal cross-lingual sentence representations. *CoRR* abs/1803.01400 (2018).
[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Annual Conference on Neural Information Processing Systems*. 5998–6008. Retrieved from https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
[25] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR* abs/1207.0580 (2012).
[26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR* abs/1910.10683 (2019).
[27] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. *CoRR* abs/2010.11934 (2020).
[28] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR* abs/1911.02116 (2019).
[29] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Retrieved from https://hal.inria.fr/hal-02131630.

Received 22 September 2022; revised 1 February 2023; accepted 13 March 2023