IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 9, 2021, accepted July 19, 2021, date of publication July 21, 2021, date of current version August 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3099021

Biomedical Text Similarity Evaluation Using Attention Mechanism and Siamese Neural Network

ZHENGGUANG LI[1], (Member, IEEE), HENG CHEN[1], (Member, IEEE), AND HUAYUE CHEN[2]

[1]Research Center for Language Intelligence, Dalian University of Foreign Languages, Dalian, Liaoning 116044, China
[2]School of Computer Science, China West Normal University, Nanchong 637002, China

Corresponding author: Huayue Chen (sunnyxiaoyue20@cwnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772103, Grant 61572102, and Grant 61702080; and in part by the Natural Science Foundation of Liaoning Province, China, under Grant 20180551078.

ABSTRACT It is a crucial component to estimate the similarity of biomedical sentence pair. Siamese neural network (SNN) can achieve better performance for non-biomedical corpora. However, SNN alone cannot obtain satisfactory biomedical text similarity evaluation results due to syntactic complexity and long sentences. In this paper, a cross self-attention (CSA) is proposed to design a new attention mechanism, namely self2self-attention(S2SA). Then the S2SA is introduced into SNN to construct a novel self2self-attentive siamese neural network, namely S2SA-SNN. In the S2SA-SNN, self-attention is used to learn the different weights of words and complex syntactic features in a single sentence. The means of the CSA are used to learn inherent interactive semantic information between sentences, and it employs self-attention instead of global attention to perform cross attention between sentences. Finally, three biomedical benchmark datasets of Pearson Correlation of 0.66 and 0.72/0.66 on DBMI and CDD-ful/-ref are used to test and prove the effectiveness of the S2SA-SNN. The experiment results show that the S2SA-SNN can achieve better performances with pre-trained word embedding and obtain better generalization ability than other compared methods.

INDEX TERMS Self-attention, cross attention, siamese network, semantic textual similarity, interactive semantic information.

I. INTRODUCTION

More and more medical texts have been accumulated with an amount of biomedical information growing. The biomedical text similarity evaluation method is a critical task in drug and drug interaction(DDI), question answering [1]–[3], etc. Although some researchers utilized biomedical resources [4] to improve the evaluation similarity performance, the generalization of these methods is poor due to the limitation of resources and corpus. Therefore, deep neural network-based methods such as character-based [5], inter-weighted alignment [6] are proposed. Furthermore, some researchers utilized word embedding [7], sentence embedding [8], [9] and shared sentence encoder [10] to obtain sentence semantic representation and estimate the similarity. Meanwhile, some researchers employed Siamese neural networks (SNN) [11] consisting of dual recurrent neural networks with shared parameters, to model sentence pairs and compute the similarity via distance function. In addition, the attention mechanisms [12] are integrated with SNN to focus on crucial words. These words have an important impact on the sentence semantic representation. These neural networks with attention mechanisms achieve good results, but they ignored the importance of interactive information between sentences. Therefore, some methods apply interactive attention [13] and cross attention [14] mechanism to obtain the interaction semantic information between sentences. The interaction contributed to enhance the semantic information of two sentences, and promise the semantic similarity estimation performance.

Even though these methods with interactive/cross attention mechanism show the effectiveness on non-biomedical

The associate editor coordinating the review of this manuscript and approving it for publication was Jesus Felez.

datasets, their performance on biomedical corpora is unsatisfactory owing to long-range dependencies [15] and complex syntactical structure [16] in biomedical corpora. Inspired by self-attention [15] and interactive attention [17], interactive self-attention has been proposed in our previous work [18]. Other methods are proposed for this field [19]–[24]. However, the semantic loss might be introduced by the semantic vector averaging operation in the interactive attention. Therefore, interactive attention is replaced by cross attention in this paper, forming a novel attention mechanism, named cross self-attention (CSA). Meanwhile, the hybrid attention mechanism based on integrating the self-attention and CSA is proposed, i.e., self2self-attention (S2SA). The S2SA mechanism is introduced into SNN to evaluate the similarity of sentence pairs and to verify the effectiveness of S2SA. The proposed attention mechanism consists of self-attention in a single sentence and CSA between sentences. Firstly, our attention mechanism learns the attention weights between words and complex syntactical features from the long/complex biomedical sentences via self-attention in a single sentence. Secondly, the attention mechanism employs CSA to obtain interactive semantic information. The interactive information is more helpful for enhancing the sentence semantic representation and alleviating the semantic loss in the interactive attention network [13].

The main contributions of this paper can be summarized as follows:

- A cross self-attention mechanism is proposed to realize semantic interaction between sentences and reduce semantic loss to a certain extent.
- A self2self-attention mechanism with composing of self-attention in a single sentence and cross self-attention between sentences, is proposed to estimate the semantic textual similarity.

II. S2SA-SNN

The semantic textual similarity estimation at sentence-level involves two sentences. Given one sentence X and another sentence Y, the goal of the proposed model is i)to learn the sentence semantic representation of X and Y, and ii) calculate a score to measure their similarity or obtain the output of Softmax activation function via the semantic representations. As shown in FIGURE 1, the model first learns basic semantic representation via the double Siamese neural network which takes biomedical word embeddings as inputs to obtain context information for each word (Sec. A). Biomedical texts are mainly collected from biomedical literature or clinical notes in this paper, these sentences are middle/long and syntactically complex. Learn long-range dependencies are a key challenge in these sentence pairs. Thus, the self-attention mechanism is introduced into our model to learn the semantic vector of each word in a sentence (Sec. B). Moreover, the researchers described the same contents/opinions using the sentences, which are consisted of the same words (synonyms, near-synonyms) with different positions in biomedical literature. Although both interactive attention [17] and cross attention [14] can learn interactive semantic information, interactive attention might introduce semantic loss owing to the semantic vector averaging. Therefore, cross self-attention is proposed in section C to obtain interactive semantic information. The hybrid attention, self2self-attention, consists of self-attention in a single sentence and CSA between sentences due to the different role of semantic information in a single sentence and interactive semantic information (Sec. D). Finally, the prediction of the proposed model is given by measuring similarity or active function (Sec. E).

A. SIAMESE NEURAL NETWORK

Siamese neural networks(SNN) consist of dual-branch networks with shared weights [11]. Therefore, they are applied to sentence/word pair tasks, such as textual similarity [25]. Moreover, bi-directional long short-term memory (bi-LSTM) has achieved good results on other biomedical NLP tasks like Named Entity Recognition (NER) [26]. Furthermore, LSTM is helpful for solving the problem of the vanishing gradient problem suffered by standard RNN in which backpropagated gradients become vanishingly small over long sequences. Hence, bi-LSTM networks are usually chosen as a branch network of SNN. However, unlike the standard SNN, each branch network is a double layer bi-LSTM network in this paper due to syntactical complexity and sentence length in biomedical texts.

Given a sentence pair $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_m)$, where n and m are the length of X and Y, respectively. Sentence X and Y are converted word embedding matrix and then separately fed into upper and lower bi-LSTM branch network. The forward and backward hidden vectors of each branch network at time-step t are described $\overrightarrow{H^t} = [\overrightarrow{h_1^t}, \overrightarrow{h_2^t}, \ldots, \overrightarrow{h_n^t}]$ and $\overleftarrow{H^t} = [\overleftarrow{h_1^t}, \overleftarrow{h_2^t}, \ldots, \overleftarrow{h_n^t}]$. Then $\overrightarrow{h_i^t}$ and $\overleftarrow{h_i^t}$ are concatenated to one vector representation, namely $h_i^t = [\overrightarrow{h_i^t}; \overleftarrow{h_i^t}]$. Finally, the hidden vector of one sentence is $H^t = [h_1^t, h_2^t, \ldots, h_n^t]$. Therefore, the output of SNN corresponding to X and Y at time-step t is described as:

$$H_x^t = [h_{x,1}^t, h_{x,2}^t, \ldots, h_{x,n}^t] \tag{1}$$

$$H_y^t = [h_{y,1}^t, h_{y,2}^t, \ldots, h_{y,n}^t] \tag{2}$$

In here, the hidden output of each LSTM cell can be calculated by equations(3)∼(8).

$$f_t = \sigma(W_f \bullet [h_{t-1}, x_t] + b_f) \tag{3}$$

$$i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i) \tag{4}$$

$$\tilde{C}_t = \tanh(W_C \bullet [h_{t-1}, x_t] + b_c) \tag{5}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{6}$$

$$o_t = \sigma(W_o \bullet [h_{t-1}, x_t] + b_o) \tag{7}$$

$$h_t = o_t * \tanh(C_t) \tag{8}$$

where, $x_t \in R^d$ is the input at time-step t, and d is the feature dimension for each word, σ is the element-wise sigmoid

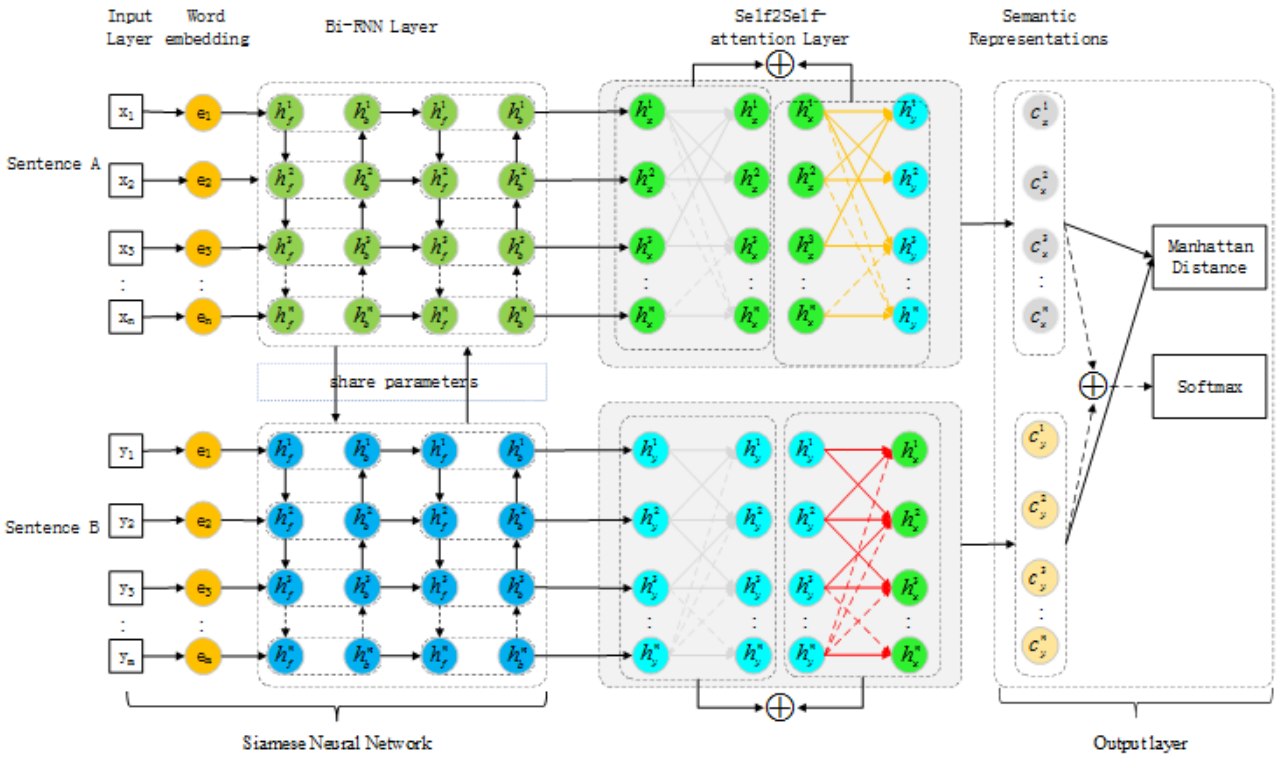

FIGURE 1. The architecture of the self2self-attentive Siamese neural network (S2SA-SNN). The model consists of siamese neural network (SNN), self2self-attention (S2SA) layer and output layer. SNN is used to learn basic semantic information of sentence pairs. S2SA is employed to capture local semantic information in a sentence and learn mutual semantic between sentences. Finally, we adopt Manhattan distance formula to evaluate the semantic similarity, and Softmax to predicate the classes.

function, • is the element-wise product. C_t is the memory cell designed to lower the risk of vanishing/exploding gradient, and therefore enabling learning of dependencies over larger period of time feasible with traditional recurrent networks. $\tilde{C}_t$ is the temporary state at time-step t. The forget gate, f_t is to reset the memory cell. i_t and o_t denote the input and output gates, and essentially control the input and output of the memory cell. tanh is the activation function.

B. SELF-ATTENTION

In fact, some parts of the sequence can be more relevant compared to others [27], namely the contribution of each word to the sentence semantic representation is different in a sentence. Therefore, some researchers proposed attention mechanisms to get different weights for denoting the different contributions to the semantic representation [28]. On the other hand, syntactic structure is relevant to the sentence semantic representation, i.e., the relationship between words with different positions has different influence on semantics. Furthermore, choosing appropriate mechanisms/methods is necessary for biomedical sentences with complex syntactic structure. Meanwhile, self-attention obtains weights of words via attention operation that is performed each word towards all words in the sentence. These weights represent the contribution of different words and syntactic structures to semantic representation. Thus, self-attention is more suitable than other attention mechanisms for biomedical sentence semantic representation.

An attention function can be regarded as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors [15]. The attention weight of each value is calculated by a compatibility function of the query with the corresponding key. Given a matrix Q, K and V denoting a set of queries, keys, and values, respectively, the output matrix of self-attention as:

$$Q, K, V = XW \tag{9}$$

$$SA(Q, K, V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V \tag{10}$$

where, $Q \in R^{n\times d_k}$, $K \in R^{m\times d_k}$, $V \in R^{m\times d_k}$, d_k is the dimension of each query q in Q and key k in K, the weight matrix W is a learning parameter, the X is an input matrix.

C. CROSS SELF-ATTENTION

Our previous work has verified the impact of semantic interaction on the semantic textual similarity estimation between sentences. Although our previous proposed interactive self-attention contributes to improving the performance owing to its semantic interaction, semantic loss might exist due to the average operation of semantic vectors. Therefore, cross self-attention (CSA) is proposed to reduce semantic

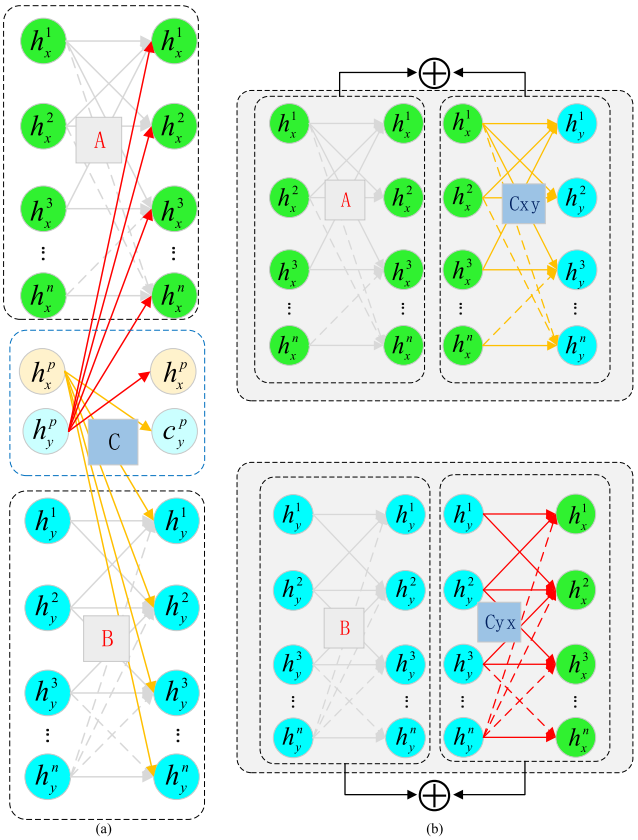

FIGURE 2. The difference between interactive self-attention and cross self-attention. (a) illustrates previous work, namely, interactive self-attention; (b) illustrates the proposed self2self-attention. Part A and B denotes self-attention in (a) and (b). Part C denotes the interaction between the mean vector and other vectors. Part Cxy and Cyx demonstrate the cross self-attention between sentences. The difference between Cxy and Cyx is the different enhanced main sentence (the enhanced main sentence of Cxy is sentence X, and that of Cyx is sentence Y).

loss. It directly adopts a similar self-attention to implement the semantic interaction, replacing the interactive attention, as shown in Figure 2. Figure 2(a) shows the basic framework of interactive self-attention, but the architecture of S2SA is shown in Figure 2(b). A and B parts in Figure 2(a) and Figure 2(b) both denote self-attention operation. C part in Figure 2(a) denotes interactive operation. Cxy and Cyx parts in Figure 2(b) denote CSA. Therefore, CSA replaces the average operation of semantic vectors with similar self-attention to reduce semantic losses.

Given the two semantic vectors $X = (x_1, x_2, \ldots x_n)$ and $Y = (y_1, y_2, \ldots y_n)$, according to the definition of self-attention, the Q, K, V of X and Y are defined as:

$$Q_X, K_X, V_X = XW_X \tag{11}$$

$$Q_Y, K_Y, V_Y = YW_Y, \tag{12}$$

where W_X and W_Y are weight matrix of X and Y respectively. Then, the cross self-attention is described as equation 13.

$$CSA(X, Y) = soft\max(\frac{Q_X K_Y^T}{\sqrt{d_X}})V_X, \tag{13}$$

where $CSA(X, Y)$ denotes the interactive semantic information from Y towards X, d_X represents the dimension of each vector X. In fact, to obtain the interactive semantic vector from Y towards X, the dimension of X and Y are equal.

To clear describe how does the cross self-attention works, an example is illustrated in Figure 3. We take a sentence pair (sentence X: "The patient was transferred to the Title."; sentence Y: "The plan was discussed with patient/family and they are in agreement.") as an example. Part (1) in Figure 3 describes the CSA operation from sentence X towards sentence Y. The semantic similarity between any word in sentence X and each word in sentence Y is calculated, formed a vector. Like this, a matrix can be obtained. The output of the Softmax operation on the matrix is used to denotes the attention weights of sentence X towards sentence Y. Part (2) shows the opposite meaning. Part (3) denotes self-attention operation of sentenc X. The semantic similarity between any word in sentence X and each word in sentence X is calculated.

D. SELF2SELF-ATTENTION

Existing methods put more emphasis on learning the separate sentence representations. For instance, Tang *et al.* [10] utilized the rich annotation data in a rich resource language to perform semantic textual similarity between sentences. Zhu *et al.* [29] proposed a dependency-based LSTM model to learn sentence representation. Their experimental results show that semantic information of a single sentence also plays an important role. The CSA is helpful for enhancing the semantic representation to each other. Meanwhile, self-attention can precisely capture semantic information of long biomedical sentence and reduce long-range dependencies problem. In other words, they are highly complementary to each other. Therefore, it is necessary to combine the advantages of self-attention and CSA. The self2self-attention, integrating self-attention with CSA, is proposed in this paper. Furthermore, to avoid semantic loss caused by the pooling of hidden states, the attention is directly applied over the final hidden state of our Siamese neural network. Finally, the hybrid attention contains i)self-attention in the X or Y, and ii)self-attention between X and Y, namely CSA. Furthermore, CSA stands for the mutual attention between X and Y. The CSA consists of two parts: the CSA of X towards Y and the CSA of Y towards X. Here, the X and Y refer to the final hidden state of the corresponding branch network in the Siamese neural network.

According to the definition in section C, the final output matrix of self2self-attention as:

$$S2SA(Q_X, K_X, V_X, K_Y)_X = \lambda SA(Q_X, K_X, V_X) + (1-\lambda)CSA(X, Y) \tag{14}$$

$$S2SA(Q_Y, K_Y, V_Y, K_X)_Y = \lambda SA(Q_Y, K_Y, V_Y) + (1-\lambda)CSA(Y, X), \tag{15}$$

where equations 14 and 15 denote the basic semantic representation of sentence X and Y, respectively. λ is a

TABLE 1. The statistics of the corpora.

	CDD-ref	CDD-ful	DBMI
Train	2051	2068	1655
Test	520	520	412
Total	2571	2588	2067

learning parameter or hyperparameter, d is the length of a sequence.

E. OUTPUT LAYER

The aforementioned outputs of self2self-attention are the final semantic representations of the sentence pair. Moreover, the sentence pair tasks are generally regarded as similarity estimation or prediction classification. Therefore, the evaluation functions in the output layer are defined as follows.

1) SIMILARITY ESTIMATION

Similarity measurement is calculated by distance functions, such as Manhattan distance formula (i.e., equation 16), Euclidean distance, cosine similarity, etc.

$$Sim(C_x, C_y) = \exp(-\left\|C_x - C_y\right\|_1), \tag{16}$$

where C_x and C_y are the outputs of the self2self-attention layer as shown in equations 14 and 15. Then the evaluation score on the test dataset is computed by evaluation functions such as Pearson, Spearman, Jaccard coefficient, etc.

2) PREDICATION CLASSIFICATION

To predicate the classes, the sentence semantic representations of sentence pairs are concatenated to form the final vector, which is then fed into a Softmax layer to predict the result as shown in equation 17.

$$\hat{y} = \underset{y \in Y(pair(x,y))}{\arg\max} \ (soft\max([C_x; C_y])), \tag{17}$$

where $\hat{y}$ is the predication class, $pair(x, y)$ denotes a sentence pair.

III. EXPERIMENTS

In this section, the proposed model named S2SA-SNN is evaluated using three biomedical datasets. Firstly, the experimental datasets and evaluation metrics are introduced. Then we describe the hyperparameters and related resources. Finally, we list the results of our model and other methods.

A. DATASETS AND EVALUATION METRICS

In our experiments, three biomedical corpora used in previous work are employed to verify the effectiveness of the proposed model and perform comparison experiments of the baselines, namely DBMI, CDD-ref and CDD-ful. The statistics of the corpora are list in TABLE 1.

Moreover, the three corpora mentioned above are converted into binary classification datasets for conducting classification experiments. Firstly, annotated scores of DBMI ([0-5]) and CDD-ful/-ref ([1-5]) are converted classification class(0 or 1). Give the middle value is $m_v = \frac{(b+a)}{2}$. The class label is 1 when the annotated score is larger than m_v, otherwise, the class label is 0, where a and b are upper and lower boundaries of annotated interval respectively.

Finally, Pearson(r), Spearman(ρ) correlation coefficient and mean square error (MSE) is used to evaluate the performance of the similarity. Meanwhile, accuracy, precision, recall, and F1-Score are adopted to evaluate the performance of a binary classification task.

TABLE 2. Experimental hyperparameters.

Hyperparameter	Value	hyperparameter	Value
epochs	50	batch size	30
learning rate	1e-5	dropout rate	0.5
bi-LSTM units	128/64	sequence length	100
optimization method	Adam		

B. HYPERPARAMETER AND RELATED RESOURCES

We implemented our models using Keras running on top of backend TensorFlow and Python3.6. Furthermore, the pre-trained biomedical word embedding can be obtained via link URL: http://evexdb.org/pmresources/ngrams/PubMed/ in website (i.e., http://bio.nlplab.org/). Mean square function and cross-entropy error function are used as the loss functions of estimation similarity and classification task individually. λ in eq. 14 and eq.15 is set 0.5 due to the interchangeability of similar textual semantic estimation task. Other hyperparameters are shown in TABLE 2.

IV. RESULTS AND ANALYSIS

A. BASELINES AND OUR MODELS

To demonstrate the effectiveness of the proposed model, we compare it against multiple baseline methods and state-of-the-art approaches for the sentence pair similarity estimation task on other corpora.

- **MaLSTM**: proposed by Mueller [25], achieving the state of the art results on SICK [30] corpus.
- **ImprovedSNN**: proposed by Chi and Zhang[31], employing hierarchical attention [28] to give different words different attention weights and achieving better results on a large dataset downloaded from Stanford Web.
- **AttentiveSNN**: proposed by Bao et.al. [32], regarding the attention weights as the coefficient of the Manhattan distance and achieving a higher Pearson correlation score than other methods on cross-lingual textual similarity corpus.
- **SNN**: our baseline like MaLSTM, but double biLSTMs in each branch network are employed in the model.
- **IA-SNN**: our baseline introducing interactive attention[17] into Siamese neural network.

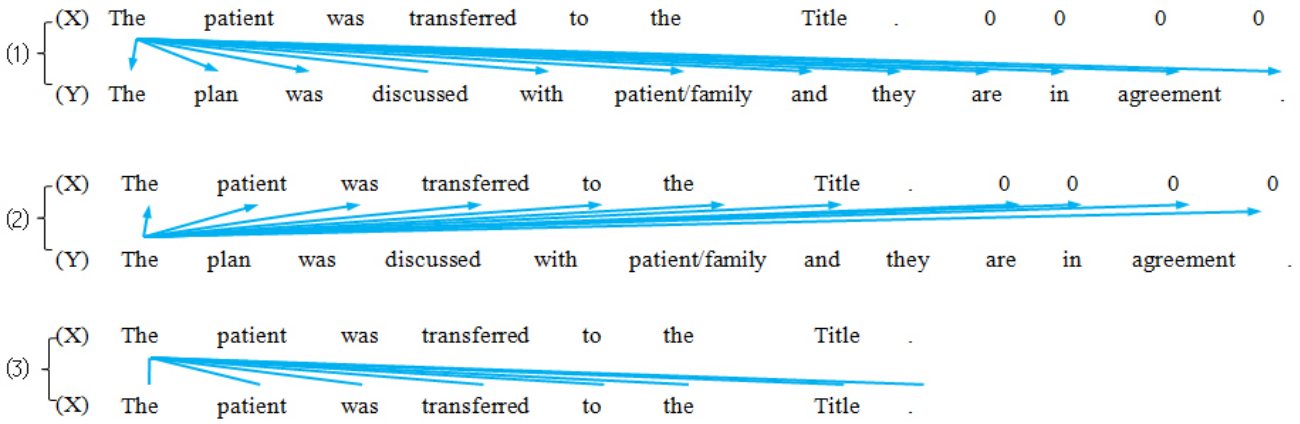

FIGURE 3. An example of CSA. Setence X is "The patient was transferred to the Title.", and sentence Y is "The plan was discussed with patient/family and they are in agreement.". (1) denotes cross self-attention operation from sentence X to sentence Y. (2) denotes cross self-attention operation from sentence Y to sentence X. (3) denotes the self-attention operation of sentence X.

TABLE 3. Performance comparison of our method and other existing methods on DBMI dataset.

	CDD-ref			CDD-ful			DBMI		
Method	r	ρ	MSE	r	ρ	MSE	r	ρ	MSE
MaLSTM[19]	0.486	0.511	1.244	0.611	0.615	1.45	0.518	0.475	2.833
ImprovedSNN[25]	0.622	0.649	0.895	0.682	0.686	1.154	0.573	0.554	2.153
AttentiveSNN[26]	0.625	0.638	0.868	0.669	0.679	1.184	0.606	0.584	1.953
IA-SNN[17]	0.651	0.654	0.873	0.653	0.65	1.255	0.612	0.567	1.734
ISA-SNN[18]	0.658	**0.664**	0.803	0.713	0.719	1.057	0.656	**0.614**	**1.512**
S2SA-SNN	**0.661**	0.662	**0.799**	**0.720**	**0.721**	**1.026**	**0.659**	0.602	1.712

- **ISA-SNN**: our previous work [18], fusing interactive attention and self-attention to implement semantic interaction between sentences and integrating it into Siamese neural network.
- **SA-SNN**: introducing self-attention into our SNN.
- **S2SA-SNN**: the proposed model, integrated self2self-attention (S2SA) with our SNN.

B. PERFORMANCE COMPARISON WITH OTHER EXISTING METHODS

To show the validity of our model, we report results on CDD-ref/-ful and DBMI corpus. The results obtained from applying our model to the test sets are shown in Table 3. This table shows both ImprovedSNN and AttentiveSNN outperform MaLSTM across all three evaluation criteria, even if they only utilize the simple attention mechanism to assign the weights of words in a sentence. This indicates that the contributions of different words are different. Thus, assigned different weights are useful for improving the performance of similarity estimation between biomedical sentences. Nevertheless, IAN obtains the weights of words via global attention and improves semantic representation by means of interactive attention. Therefore, compared with AttentiveSNN, the MSE score of IAN increases by 5%, and it achieves the best Spearman correlation coefficient on CDD-ref. Moreover, the overall performance is better than the other three methods on DBMI while there is a decrease on CDD-ful. Furthermore, IAN attains the worst performance in all the methods with the attention mechanism on CDD-ful. This shows that the performance of IAN may depend on the quality of the corpus and the complexity of the sentences in the datasets. The analysis of the corpora will be given later. The ISA-SNN outperforms all the methods on the three corpora owing to alleviating long-range dependencies by self-attention. Finally, the proposed model achieves the best results by the self2self-attention on the three datasets except for the Spearman correlation coefficient on CDD-ref. The final Pearson correlation coefficient is increased to 0.661, approaching the official value 0.678 on CDD-ref[33]. The reasons for the better performance obtained by our model on the three corpora may be that the proposed self2self-attention not only helps our model to more precisely represent the semantic via self-attention in a single sentence, but

TABLE 4. The effect of self-attention and cross self-attention on CDD-ful/-ref and DBMI.

	CDD-ref			CDD-ful			DBMI		
Method	r	p	MSE	R	p	MSE	r	p	MSE
SNN	0.531	0.545	1.161	0.511	0.518	2.287	0.458	0.466	3.296
SNN +Self-attention(SA-SNN)	0.575	0.619	1.128	0.678	0.683	1.187	0.554	0.527	2.113
SNN+Self-attention +CSA(S2SA-SNN)	**0.661**	**0.662**	**0.799**	**0.72**	**0.721**	**1.026**	**0.659**	**0.602**	**1.712**

enhance the sentence semantic representation through cross self-attention. Moreover, the results of the proposed model are higher than that of the previous proposed model. This shows that the vector averaging in IAN causes semantic loss and the CSA is useful for reducing the effect of the problem.

C. THE EFFECT OF SELF-ATTENTION AND CROSS SELF-ATTENTION

We also investigate the effect of self-attention and cross self-attention on the performance in our model. Table 4 shows the results of the two attention mechanisms on the CDD-ful/-ref and DBMI. The baseline is our Siamese neural networks(SNN), whose each branch network contains dual layers bi-LSTM. Then, self-attention is introduced into our SNN(named SA-SNN). Finally, cross self-attention (CSA) is integrated with self-attention, named self2self-attention. Moreover, the self2self-attention is added into SNN(named S2SA-SNN).

Firstly, three evaluation scores on the three datasets are improved significantly. Moreover, the Pearson and Spearman correlation coefficient is increased by 0.16 and 0.17, respectively on CDD-ref. This excellent performance may benefit from the precise semantic representation, i.e., weights of words and syntactic structure learned by self-attention within the query and answer. Secondly, on DBMI, CDD-ref, CDD-ful datasets, the cross self-attention yields a boost of up to 0.1, 0.09, 0.06 Pearson correlation coefficient over self-attention separately. Therefore, the sentence enhancement semantic representation gained by cross self-attention between the query and the answer is useful for improving performance. Although S2SA-SNN achieves the best results on CDD-ref, its increase based on self-attention is the smallest one. On the contrary, the increase of SA-SNN in the Pearson and Spearman correlation coefficient based on the baseline is the largest. Therefore, an investigation into the difference among the three datasets is opened in the following part of this section.

The maximum difference is the number of long and other sentence pairs. As shown in FIGURE 4, the number of sentence pairs with more than 80 words in CDD-ful is more than that in the CDD-ref and DBMI (20.6%, 10.7%, 4.6%, respectively). In addition, more special tokens like ''Figure'', ''()'' are found in CDD-ful. This indicates that i) self-attention

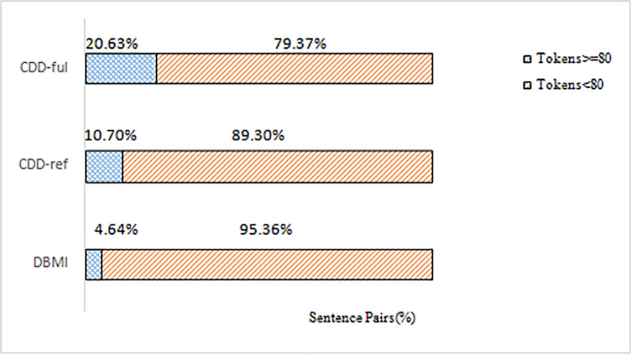

FIGURE 4. Statistics on the number of tokens within a sentence in the three datasets.

is more effective than other attention mechanisms to estimate the similarity between the long and complex sentences, and ii) although cross self-attention is helpful to improve the performance on different corpora, it is more effective for the improvement of short and medium sentences. Furthermore, S2SA-SNN also promotes the results of corpora with a small amount of noise. To further illustrate the performance of the proposed method on long sentence pairs, the test sets are divided into long and other sentence pairs and recalculated the evaluation scores as shown in Table 5. This table shows that the improvement scope of SA-SNN is larger than that of the other three attentive methods on long sentence pairs of CDD-ful, and S2SA-SNN achieves the best results. On the contrary, S2SA-SNN outperforms the other methods while the performance of ImprovedSNN exceeds that of the SA-SNN on short/medium sentence pairs. Meanwhile, the results obtained by S2SA-SNN is better than other methods on the long sentence pairs of CDD-ref, but the results of SA-SNN is worst relative to other methods. Furthermore, both SA-SNN and S2SA-SNN don't attain the best results on the DBMI corpus.

D. PERFORMANCE COMPARISON OF OUR METHOD AND BASELINE WHEN REGARDING AS BINARY CLASSIFICATION TASK

To investigate the ability of classification on sentence pairs, the experiments with the converted binary classification datasets are conducted using SNN, SA-SNN, and S2SA-SNN. The performance of tasks as mentioned above is

TABLE 5. Performance comparison of different attentive methods on long/short sentence pairs.

		CDD-ref			CDD-ful			DBMI		
Method		r	p	MSE	r	p	MSE	r	p	MSE
Long	ImprovedSNN[25]	0.601	0.616	0.937	0.592	0.599	1.305	0.673	0.72	0.358
	AttentiveSNN[26]	0.577	0.539	0.809	0.601	0.609	1.314	0.648	0.62	0.305
	IAN[17]	0.659	0.637	0.697	0.561	0.564	1.442	**0.72**	**0.784**	**0.282**
	SA-SNN	0.542	0.55	1.269	0.659	**0.665**	1.168	0.659	0.658	0.375
	S2SA-SNN	**0.692**	**0.657**	**0.623**	**0.667**	0.664	**1.103**	0.657	0.662	0.334
Short	ImprovedSNN	0.657	0.61	**0.691**	0.711	0.714	1.098	**0.64**	**0.586**	1.824
	AttentiveSNN	0.632	0.655	0.88	0.692	0.703	1.136	0.619	0.557	**1.542**
	IAN	**0.657**	**0.668**	0.824	0.683	0.679	1.186	0.625	0.573	1.74
	SA-SNN	0.585	0.636	1.101	0.684	0.687	1.194	0.581	0.532	1.909
	S2SA-SNN	0.656	0.661	0.834	**0.737**	**0.737**	**0.998**	0.611	0.565	1.697

TABLE 6. Performance comparison of our method and baseline when regarding as binary classification task.

	CDD-ful/-ref				DBMI			
Method	P	R	F1-score	Accuracy(%)	P	R	F1-score	Accuracy(%)
SNN	**0. 81**/0.654	0.33 /**0.744**	0. 469/0.696	62.9 /69.5	0.3957	0.875	0.5450	54.50
SA-SNN	0.804/0.658	0.722/0.651	0.761/0.654	77.5/76.3	**0.6**	0.847	**0.702**	**79.5**
S2SA-SNN	0.804/**0.737**	**0.808**/0.523	**0.806**/0.612	**80.7**/**77.1**	0.575	**0.894**	0.7	78.1

shown in Table 6. The SNN achieves the best precision rate on CDD-ful, but the worst F1-score and accuracy on the three datasets. However, the recall rate, F1-score, and accuracy of the S2SA-SNN are higher than that of SNN and SA-SNN on CDD-ful. The results reveal that applying self2self-attention in corpora with main long and syntactic complex sentences has some advantages over methods without attention and self-attention due to the weights and complex syntactic features learned by self-attention and interactive semantic information obtained cross self-attention between sentences. However, the recall rate of the S2SA-SNN outperforms that of the SA-SNN while the precision of the SA-SNN is better than S2SA-SNN on DBMI. Meanwhile, S2SA-SNN only attains the best precision rate and accuracy relative to other methods. Therefore, the overall classification results achieved by S2SA-SNN are worse than SA-SNN on short/medium sentences pairs. The reason may be that the introduction of cross self-attention causes a small amount of noise that has an impact on the classification performance.

In addition, to analyze the generalization performance of three methods, we draw ROC curve and compute AUC score on CDD-ful dataset, as shown in Figure 5. It demonstrates that the AUC of the S2SA-SNN(0.90) outperforms the SNN without attention (SNN: 0.80) and the SA-SNN (introduced self-attention, 0.86) in the classification task. Thus, The S2SA-SNN is more suitable for applying to other sentence pair datasets with long or complex syntactic sentences.

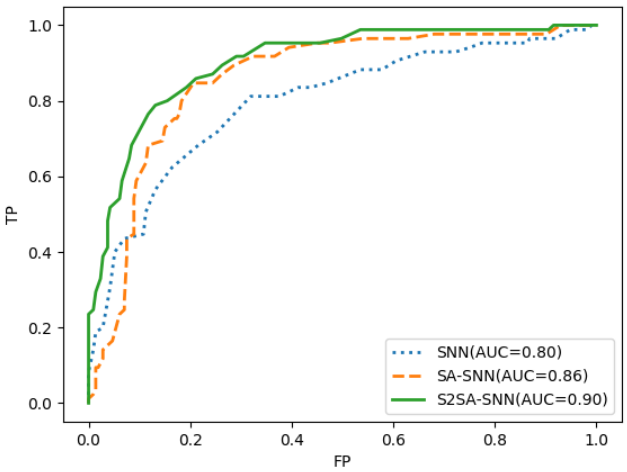

FIGURE 5. ROC and AUC evaluation of SNN,SA-SNN and S2SA-SNN ON CDD-ful DATASET. TP denotes true positive rate, and FP denotes false positive rate.

E. ADVANTAGES AND LIMITATION OF THE PROPOSED MODEL

To analyze the computational efficiency, the mean running time, CPU/GPU occupancy of each epoch of SNN, SA-SNN, ISA-SNN, S2SA-SNN are shown in Table 7. Firstly, there is no obvious difference in the efficiency of the six methods in terms of time, only a few milliseconds. S2SA-SNN took three milliseconds more than SNN evaluation. Moreover, in terms

TABLE 7. Computational efficiency comparison on DBMI corpus.

Methods	Time(s)	CPU(%)	GPU(%)
SNN	12.017	8.3-9.5	56-60
SA-SNN	12.019	8.7-9.3	53-56
IA-SNN	12.017	8.2-8.7	52-59
ISA-SNN	12.020	7.8-8.5	52-54
S2SA-SNN	12.019	8.1-8.9	54-56

of CPU and GPU occupancy, there is almost no difference. The occupancy of SNN in GPU and CPU is slightly higher than that of S2SA-SNN.

Combined with results of short / long sentences and binary classification and computational efficiency analysis, S2SA owns the following advantages: i) it is more suitable for sentence pair corpus with long or complex syntactic sentences. ii) The generalization of S2SA is better than SNN and SA-SNN. iii) The computational efficiency is not lower than other methods. However, it also has some limitations. First, its performance is not better than simple SNN on the datasets with more short sentences. Second, it is sensitive to noisy data, thus, it is not recommended to be applied in datasets with noise texts.

V. CONCLUSION AND FUTURE WORK

In this paper, a cross self-attention is proposed, which is integrated with self-attention for designing a novel hybrid attention mechanism, namely self2self-attention mechanism. Finally, the proposed hybrid attention is introduced into the Siamese neural network with bidirectional LSTM, called self2self-attentive Siamese neural network (S2SA-SNN). It can represent the sentence semantic more precisely in a single sentence via self-attention on basis of shared parameters of the Siamese network. Moreover, inherent interactive semantic information between sentences is learned via the cross self-attention. The semantic loss is alleviated by CSA owing to removing vector averaging operation. Consequently, the interactive information learned by CSA contributes to enhancing the sentence semantic representation and improving the overall performance. Furthermore, we conduct experiments on three biomedical datasets. Experimental results indicate that the proposed model for measuring biomedical textual similarity and classifying sentence pairs has a better performance on the three datasets. The analyses of long / short sentences and corpus indicate that self2self-attention is more suitable for datasets with long or complex syntactic and less noise sentences. Our model depends on traditional context-independent word embeddings only to verify the effectiveness of cross self-attention and self2self-attention. In addition, we can combine external biomedical knowledge into our model.

REFERENCES

[1] L. He, Z. Yang, Z. Zhao, H. Lin, and Y. Li, "Extracting drug-drug interaction from the biomedical literature using a stacked generalization-based approach," *PLoS ONE*, vol. 8, no. 6, Jun. 2013, Art. no. e65814.
[2] L. Cai, S. Zhou, X. Yan, and R. Yuan, "A stacked BiLSTM neural network based on coattention mechanism for question answering," *Comput. Intell. Neurosci.*, vol. 2019, pp. 1–12, Aug. 2019, doi: 10.1155/2019/9543490.
[3] W. Deng, S. Shang, X. Cai, H. Zhao, Y. Zhou, H. Chen, and W. Deng, "Quantum differential evolution with cooperative coevolution framework and hybrid mutation strategy for large scale optimization," *Knowl.-Based Syst.*, vol. 224, Jul. 2021, Art. no. 107080.
[4] M. B. Aouicha and M. A. H. Taieb, "Computing semantic similarity between biomedical concepts using new information content approach," *J. Biomed. Informat.*, vol. 59, pp. 258–275, Feb. 2016.
[5] W. Lan and W. Xu, "Character-based neural networks for sentence pair modeling," 2018, *arXiv:1805.08297*. [Online]. Available: https://arxiv.org/abs/1805.08297
[6] G. Shen, Y. Yang, and Z.-H. Deng, "Inter-weighted alignment network for sentence pair modeling," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 1179–1189.
[7] T. Kenter and M. de Rijke, "Short text similarity with word embeddings," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2015, pp. 1411–1420, doi: 10.1145/2806416.2806475.
[8] K. Blagec, H. Xu, A. Agibetov, and M. Samwald, "Neural sentence embedding models for semantic similarity estimation in the biomedical domain," *BMC Bioinf.*, vol. 20, no. 1, pp. 1–10, Dec. 2019, doi: 10.1186/s12859-019-2789-2.
[9] W. Deng, S. Shang, X. Cai, H. Zhao, Y. Song, and J. Xu, "An improved differential evolution algorithm and its application in optimization problem," *Soft Comput.*, vol. 25, no. 7, pp. 5277–5298, Apr. 2021.
[10] X. Tang, S. Cheng, L. Do, Z. Min, F. Ji, H. Yu, J. Zhang, and H. Chen, "Improving multilingual semantic textual similarity with shared sentence encoder for low-resource languages," 2018, *arXiv:1810.08740*. [Online]. Available: https://arxiv.org/abs/1810.08740
[11] P. Neculoiu, M. Versteegh, and M. Rotaru, "Learning text similarity with siamese recurrent networks," in *Proc. 1st Workshop Represent. Learn. NLP*, Jan. 2016, pp. 148–157, doi: 10.18653/v1/W16-1617.
[12] C. Tan, F. Wei, W. Wang, W. Lv, and M. Zhou, "Multiway attention networks for modeling sentence Pairs," in *Proc. IJCAI*, 2018, pp. 4411–4417, doi: 10.24963/ijcai.2018/613.
[13] Q. Yin, G. Luo, X. Zhu, Q. Hu, and O. Wu, "Semi-interactive attention network for answer understanding in reverse-QA," 2019, *arXiv:1901.03788*. [Online]. Available: https://arxiv.org/abs/1901.03788
[14] Y. Hao, Y. Zhang, K. Liu, S. He, Z. Liu, H. Wu, and J. Zhao, "An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, Jul. 2017, pp. 221–231, doi: 10.18653/v1/P17-1021.
[15] S. Duan and H. Zhao, "Attention is all you need for Chinese word segmentation," 2019, *arXiv:1910.14537*. [Online]. Available: https://arxiv.org/abs/1910.14537
[16] H. Jang, J. Lim, J. H. Lim, S. J. Park, K. C. Lee, and S. H. Park, "Finding the evidence for protein-protein interactions from PubMed abstracts," *Bioinformatics*, vol. 22, no. 14, pp. 220–226, 2006.
[17] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," 2017, *arXiv:1709.00893*. [Online]. Available: https://arxiv.org/abs/1709.00893
[18] Z. Li, H. Lin, W. Zheng, M. M. Tadesse, Z. Yang, and J. Wang, "Interactive self-attentive siamese network for biomedical sentence similarity," *IEEE Access*, vol. 8, pp. 84093–84104, 2020.
[19] W. Deng, J. Xu, X.-Z. Gao, and H. Zhao, "An enhanced MSIQDE algorithm with novel multiple strategies for global optimization problems," *IEEE Trans. Syst., Man, Cybern., Syst.*, early access, Nov. 4, 2020, doi: 10.1109/TSMC.2020.3030792.
[20] Y. Song, D. Wu, A. W. Mohamed, X. Zhou, B. Zhang, and W. Deng, "Enhanced success history adaptive DE for parameter optimization of photovoltaic models," *Complexity*, vol. 2021, pp. 1–22, Jan. 2021.
[21] X. Cai, H. Zhao, S. Shang, Y. Zhou, W. Deng, H. Chen, and W. Deng, "An improved quantum-inspired cooperative co-evolution algorithm with muli-strategy and its application," *Expert Syst. Appl.*, vol. 171, Jun. 2021, Art. no. 114629.
[22] H. Cui, Y. Guan, H. Chen, and W. Deng, "A novel advancing signal processing method based on coupled multi-stable stochastic resonance for fault detection," *Appl. Sci.*, vol. 11, no. 12, p. 5385, Jun. 2021.
[23] Y. Song, D. Wu, W. Deng, X.-Z. Gao, T. Li, B. Zhang, and Y. Li, "MPPCEDE: Multi-population parallel co-evolutionary differential evolution for parameter optimization," *Energy Convers. Manage.*, vol. 228, Jan. 2021, Art. no. 113661.

[24] W. Deng, J. Xu, H. Zhao, and Y. Song, "A novel gate resource allocation method using improved PSO-based QEA," *IEEE Trans. Intell. Transp. Syst.*, early access, Oct. 1, 2020, doi: 10.1109/TITS.2020.3025796.
[25] J. Mueller and A. Thyagarajan, "Siamese recurrent architectures for learning sentence similarity," in *Proc. AAAI Conf. Artif. Intell.*, vol. 16, Mar. 2016, pp. 2786–2792.
[26] N. Limsopatham and N. Collier, "Learning orthographic features in bi-directional LSTM for biomedical named entity recognition," in *Proc. 5th Workshop Building Evaluating Resour. Biomed. Text Mining (BioTxtM)*, 2016, pp. 10–19.
[27] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," 2019, *arXiv:1904.02874*. [Online]. Available: https://arxiv.org/abs/1904.02874
[28] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 1480–1489.
[29] W. Zhu, T. Yao, J. Ni, B. Wei, and Z. Lu, "Dependency-based Siamese long short-term memory network for learning sentence representations," *PLoS ONE*, vol. 13, no. 3, pp. 1–14, 2018, doi: 10.1371/journal.pone.0193919.
[30] L. Bentivogli, R. Bernardi, M. Marelli, S. Menini, M. Baroni, and R. Zamparelli, "SICK through the SemEval glasses. Lesson learned from the evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment," *Lang. Resour. Eval.*, vol. 50, no. 1, pp. 95–124, Mar. 2016.
[31] Z. Chi and B. Zhang, "A sentence similarity estimation method based on improved siamese network," *J. Intell. Learn. Syst. Appl.*, vol. 10, no. 4, pp. 121–134, 2018.
[32] W. Bao, W. Bao, J. Du, Y. Yang, and X. Zhao, "Attentive Siamese LSTM network for semantic textual similarity measure," in *Proc. Int. Conf. Asian Lang. Process. (IALP)*, Nov. 2018, pp. 312–317.
[33] R. Islamaj, W. J. Wilbur, N. Xie, N. R. Gonzales, N. Thanki, R. Yamashita, C. Zheng, A. Marchler-Bauer, and Z. Lu, "PubMed text similarity model and its application to curation efforts in the conserved domain database," *Database*, vol. 2019, pp. 1–13, Jan. 2019, doi: 10.1093/database/baz064.

ZHENGGUANG LI (Member, IEEE) received the M.S. degree in computer application technology from Dalian Jiaotong University, in 2007, and the Ph.D. degree in computer science and technology from Dalian University of Technology. He is currently a Lecturer with Dalian University of Foreign Languages, Dalian, China. His research interests include text mining and natural language processing in social media and biomedical fields.

HENG CHEN (Member, IEEE) received the M.S. degree in computer application technology from Dalian Jiaotong University, in 2007. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Dalian Maritime University. He has been an Associate Professor with Dalian University of Foreign Languages, Dalian, China. His research interest includes natural language processing.

HUAYUE CHEN received the B.S. degree in computer science and technology from Sichuan Normal College, Nanchong, China, in 2002, and the M.S. degree in computer software and theory from Chongqing University, in 2005. Since 2012, she has been an Associate Professor with China West Normal University, Nanchong. Her research interests include artificial intelligence, optimization method, and image processing.