Interactive Self-Attentive Siamese Network for Biomedical Sentence Similarity

ZHENGGUANG LI¹,², HONGFEI LIN¹, WEI ZHENG², MICHAEL M. TADESSE¹, ZHIHAO YANG¹, AND JIAN WANG¹

¹College of Computer Science and Technology, Dalian University of Technology, Dalian 116023, China
²Software Institute, Dalian Jiaotong University, Dalian 116028, China

Corresponding author: Hongfei Lin (hongfeilin@dlut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61772103, Grant 61572102, Grant 61702080, and Grant 61771087, in part by the Natural Science Foundation of Liaoning Provincial, China, under Grant 20180551078, and in part by the Open Project Foundation of Intelligent Information Processing Key Laboratory of Shanxi Province under Grant CICIP2018006.

ABSTRACT The determination of semantic similarity between sentences is an important component in natural language processing (NLP) tasks such as text retrieval and text summarization. Many approaches have been proposed for estimating sentence similarity, and Siamese neural networks (SNN) provide a better approach. However, the sentence semantic representation, generated by sharing weights in the SNN without any attention mechanism, ignores the different contributions of different words to the overall sentence semantics. Furthermore, the attention operation within only a single sentence neglects interactive semantic influence on similarity estimation. To address these issues, an interactive self-attention (ISA) mechanism is proposed in this paper and integrated with an SNN, named an interactive self-attentive Siamese neural network (ISA-SNN) which is used to verify the effectiveness of ISA. The proposed model obtains the weights of words in a single sentence by means of self-attention and extracts inherent interactive semantic information between sentences via interactive attention to enhance sentence semantic representation. It achieves better performances without feature engineering than other existing methods on three biomedical benchmark datasets (a Pearson correlation coefficient of 0.656 and 0.713/0.658 on DBMI and CDD-ful/-ref, respectively).

INDEX TERMS Interactive self-attention, Siamese network, sentence pair, semantic similarity.

I. INTRODUCTION

Increasing numbers of medical texts have been accumulated with a growing amount of biomedical information. However, many sentences represent similar semantic meaning, but consist of completely different text descriptions in these considerable data, resulting in considerable unnecessary trouble for medical research. Therefore, evaluating the textual similarity between biomedical texts is an important task of extracting useful biomedical information, such as drug-drug interactions (DDIs) [1]. Although some researchers have utilized biomedical resources [2] or corpora [3] to improve the performance of evaluation similarity, the generalization of these methods is poor due to limitations in the resources and corpora. Therefore, machine learning-based methods [4] are also proposed for this task.

Some researchers have utilized word embedding [5], sentence embedding [6], [7] and shared sentence encoders [8] to obtain semantic expression and estimate similarity. Neculoiu et al. [9] and Mueller [10] computed the similarity between texts via Siamese neural networks (SNNs), containing dual recurrent neural networks. Although Siamese neural networks perform well in evaluating text similarity tasks, researchers have proposed improved models based on the standard SNN for this task. Zhu et al. [11] presented a replicated Siamese LSTM model for estimating similarity in asymmetric text and achieved better performance. Then, a dependency-based Siamese LSTM network was proposed by Zhu for sentence similarity. The model captured richer semantic information about a sentence than the standard LSTM model and learned an efficient sentence representation. However, these improved SNNs are mainly based on the improvement of a single network structure such as hierarchical LSTM, or the introduction of other mechanisms.
such as dependency. Thus, Pontes [12] extracted crucial semantic vectors via two convolutional neural networks (CNNs), and then fed the semantic vectors into the following Siamese network. This combination of networks helped to preserve the relevant information of sentences and promoted the calculation results of the similarity between sentences. Nevertheless, these methods have ignored the influence of different words on the overall semantics. An attention mechanism can calculate the weight of each word, representing the influence of words on sentence semantics. The attentive Siamese LSTM network, instead of using numerous manual features relying on prior knowledge or external resources, is presented for measuring semantic textual similarity [13].

It is insufficient to estimate the similarity between sentences via calculating the similarity/distance of two separate sentence semantic representations, which are obtained from deep neural network models. In fact, different words have different contributions to the overall semantics in a single sentence. Moreover, the semantics of one sentence also impacts the semantics of other sentences in the task of evaluating the similarity of two sentences. Inspired by self-attention [14] and interactive attention [15], interactive self-attention (ISA, integrating self-attention with interactive attention) is proposed in this paper. Then, ISA is integrated with a Siamese neural network, named the interactive self-attentive Siamese neural network (ISA-SNN). The model learns the weights of words in a single sentence by using a self-attention mechanism and sharing parameters of the SNN and captures the interactive semantic information between two sentences via an interactive attention mechanism.

The main contributions can be summarized as follows:

- An interactive self-attention (ISA) mechanism is proposed that integrates self-attention with interactive attention.
- The proposed model utilizes ISA to obtain the interactive semantic information between the biomedical sentences, and the interactive semantic information may enhance the sentence semantic representation learned by self-attention.
- Three biomedical datasets are employed to verify the effectiveness of the proposed mechanism and model. Analysis of Siamese networks with different mechanisms shows that the ISA is helpful in enhancing semantic information on the sentence pair and results in improving the performance of estimating textual similarity.

The rest of the paper is organized as follows. Section II gives a brief review of related work. Section III presents the different attention mechanisms and ISA-SNN for evaluating the similarity between sentences. Section IV describes the experiment detail, parameter settings, and evaluation metrics. Detailed analysis and a summary of the experimental results are explained in section V. In section VI, a conclusion is drawn, and the future work is discussed.

II. RELATED WORK
Evaluating textual similarity as one of the main tasks in NLP refers to the calculation of the similarity between term pairs or sentence pairs. Researchers evaluated the textual similarity using information retrieval methods [16] and automatic text summarization [17]. To overcome the difficult operation and common error of information retrieval methods, an automatic evaluation algorithm based on word cooccurrence [18] was proposed, but common words in sentences affected the scores. Therefore, knowledge- [19] and corpus-based methods were applied to measure the semantic similarity [3] and decrease the error rate. However, the performance of these methods depended on the quality of the knowledge base/corpus. Hence, such machine learning-based methods as SVM [4] were used for the task. Specifically, neural network methods such as DBN [20] and broad learning [21] achieve better performance. A Siamese network, sharing the parameters between two RNNs or other neural network models [9] was better than one. The method mainly focused on the semantics of words or a single sentence without considering the contribution of different words to semantics until an attention mechanism was used in NLP tasks.

In some NLP tasks, some parts of the input can be more relevant compared to others [22]. Thus, Bahdanau et al. [23] first introduced the attention mechanism for machine translation. Researchers used the mechanism and improved attention methods in such different NLP tasks as text classification [24], Name Entity Recognition (NER) [25], and sentiment classification [26]. First, for a single sequence, relative to soft attention [23], a hard attention model [27] in which the context vector is computed from stochastically sampled hidden states in the input sequence was used, while local and global attention [28] were used to capture the local and global context. For capturing the relationship within a sequence, self-attention [29], also known as inner attention, was proposed. Second, for multiple sequences, coattention [30] was joined with multiple source information attention to guide each other, and interactive attention [15] well represented a target and its collective context, which contributes to classification tasks. The interactive attention mechanism in such sentence/word pair tasks as question-answer is effective due to its interactive semantics [15]. Estimating the similarity between sentences is a typical sentence pair task. Although some researchers improved the Siamese network [31] and utilized different deep neural networks [8] to obtain better results, and interactive attention and self-attention were not integrated with other deep neural network-based methods for calculating the similarity between sentences. Therefore, we introduce self-attention to obtain weights of words and interactive attention to enhance semantic representation.

III. METHODS
As shown in Figure 1, the proposed model first utilizes the Siamese neural network consisting of an input layer, a word
embedding layer and dual bi-RNN layer to extract the basic semantic information of sentence pairs. Then, interactive self-attention obtains the final sentence semantic representation. Finally, a distance measuring function such as Manhattan is employed to estimate the semantic similarity between sentences.

A. SIAMESE NEURAL NETWORK

Siamese neural networks (SNNs) [9], [10] are used to evaluate the similarity between texts. They are dual-branch networks with shared weights, which consist of the same network copied and merged with an energy function [9]. Furthermore, bidirectional recurrent neural networks (bi-RNNs) have achieved good results on other biomedical NLP tasks such as named entity recognition (NER) [32]. Thus, the Siamese network employs bi-RNN as the branch networks in the paper. In addition, double bi-LSTMs in each branch network are employed for textual similarity estimation owing to syntactical complexity in biomedical texts [34]. LSTM and GRU are the two most widely used network cells of RNN. Both GRU and LSTM contain an update gate z to determine how much of the memory of a previous cell to keep and a reset gate r to determine how to combine the input of a new cell with previous memory. Nevertheless, a slight difference exists between LSTM and GRU. First, an LSTM cell has three gates, but there are only two gates in a GRU. Second, the forget and input gates of a GRU cell are combined into a single update gate z, and reset gate r is directly applied to the previous hidden state [33]. In this work, our experimental datasets consist of long and syntactically complex sentences. Hence, we use LSTM cells as units of our bi-RNN owing to capturing more complex patterns via the extra gating function [34]. The hidden output of each LSTM cell can be calculated via equations 1-6.

\[
\begin{align*}
    f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\
    i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\
    C_t &= \tanh(W_C \cdot [h_{t-1}, x_t] + b_c) \\
    C_t &= f_t \ast C_{t-1} + i_t \ast \tilde{C}_t \\
    o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\
    h_t &= o_t \ast \tanh(C_t)
\end{align*}
\]

where \(W_i, W_f, W_o \in \mathbb{R}^{d \times d}\) are the weighted matrices and \(b_i, b_f, b_o \in \mathbb{R}^{d}\) are biases of LSTM to be learned during training, \(x_t \in \mathbb{R}^{d}\) is the input at time-step t, and d the feature dimensions.
dimension for each word, $\sigma$ the elementwise sigmoid function, and $\bullet$ the elementwise product. $C_t$ is the memory cell designed to lower the risk of vanishing/exploding gradient, and therefore, enables learning of dependencies over a larger period of time than what is feasible with traditional recurrent networks. $C_t$ is the temporary state at time-step $t$. The forget gate, $f_t$, resets the memory cell. $i_t$ and $o_t$ denote the input and output gates, respectively, and essentially control the input and output of the memory cell. tanh is the activation function.

### B. SELF-ATTENTION

SNN ignores different contributions of different words to the overall sentence semantic representation while it extracts the semantics through sharing parameters. Therefore, some researchers utilized attention mechanisms to learn the different weights in a sentence [13], [35]. Additionally, the semantic similarity between words is useful for biomedical sentence similarity estimation because these sentences obtained from biomedical databases, such as PubMed and the Conserved Domain Database, are syntactically complex, and it is difficult to obtain significant semantic features. Furthermore, self-attention obtains weights of words via an attention operation that is performed between each word and all words in the sentence. These attention weights not only denote the different contributions to the overall sentence semantics but also denote the similarity between words in a sentence. Therefore, self-attention is more suitable than other attention mechanisms for biomedical sentence similarity estimation.

Given an input sequence $s = (x_1, x_2, \ldots, x_n)$, self-attention is defined as calculating the weight $\alpha_i$ of each word $x_i$ towards the other words in the sequence $s$. The weight vector $\alpha_i$ represents the contribution of word $x_i$ on the semantics of the sequences. The weight matrix of the sequences can be described as $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$, where $n$ is the number of tokens in the sentence and $\alpha_i$ is calculated by the following equation 7-10.

\[
q_i, k_i, v_i = x_i W
\]
\[
\alpha_{i,j} = q_i \bullet k_j / \sqrt{d}
\]
\[
\hat{\alpha}_{i,j} = \exp(\alpha_{i,j}) / \sum_{j=1}^{n} \exp(\alpha_{i,j})
\]
\[
\alpha_i = \frac{\sum_{j=1}^{n} \hat{\alpha}_{i,j}}{n}
\]

where $d$ is the dimension of $q$ and $k$, the weight matrix $W$ is a parameter of the model, and $\bullet$ is the elementwise product. $q_i$, $k_i$ and $v_i$ denote the vector of the $i$th word. $\alpha_{i,j}$ and $\hat{\alpha}_{i,j}$ are values of the $i$th and $j$th word using the elementwise product and softmax function, respectively. Finally, the output of self-attention is $C = (c_1, c_2, \ldots, c_n)$, where $c_i = \alpha_i v_i$ which represents the enhanced semantic vector of each word in the sentence $s$.

### C. INTERACTIVE SELF-ATTENTION

However, it is insufficient to only focus on semantic information in a single sentence for computing semantic similarity between sentences. In biomedical literature, the researchers described the same contents/opinions using sentences that consist of the same words (synonyms, near-synonym) with different positions, e.g., “Kalirin and Trio are encoded by separate genes in mammals and by a single one in invertebrates.” And “In mammals, Kalirin and Trio are encoded by separate genes, but invertebrates have a single homologous gene.” Therefore, semantic information in one sentence contributes to semantic representation in another sentence in biomedical sentence pairs. Inspired by the interactive attention network (IAN) [15], one sentence and another sentence are regarded as the context and target, respectively, in IAN. In this way, interactive attention can be used to compute the semantic similarity in the sentence pairs naturally. In this paper, self-attention is integrated with interactive attention, named interactive self-attention (ISA). The ISA extracts sentence semantics via self-attention and enhances the semantic representation through interactive attention.

Given the sequence $X = (x_1, x_2, \ldots, x_n)$ and another sequence $Y = (y_1, y_2, \ldots, y_n)$, where $X$ and $Y$ are the two semantic features of the output of the Siamese network, respectively, and $n$ is the length of the sequence. Since there are two inputs, each input has a function in the interactive self-attention network, defined as equations 11 and 12.

\[
Q^x, K^x = [X; \tilde{x}] W_x
\]
\[
V^x = X W_x
\]
\[
Q^y, K^y = [Y; \tilde{y}] W_y
\]
\[
V^y = Y W_y
\]

where $W_x$ and $W_y$ are weight matrices of the A and B sequences, respectively. Moreover, $\tilde{x}$ and $\tilde{y}$ are the average hidden states of all elements in the sequence $X$ and $Y$; is the concatenation operator. The attention weight vectors of the two sequences are described as $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_n, \beta_i)$. Here, $\alpha_1, \alpha_2, \beta_1$, and $\beta_2$ are calculated according to equation 8 - 10. Then, $\alpha_1$ and $\beta_2$ are both removed from matrices $\alpha$ and $\beta$. Finally, the sequence $X$ and $Y$ are multiplied by their respective attention weight vectors as the output of interactive self-attention, described as follows.

\[
C_x = (\alpha_1 x_1^T, \alpha_2 x_2^T, \ldots, \alpha_n x_n^T)
\]
\[
C_y = (\beta_1 y_1^T, \beta_2 y_2^T, \ldots, \beta_n y_n^T)
\]

### D. SIMILARITY ESTIMATION

In this section, the final process of similarity estimation is presented. The training set for a Siamese neural network consists of triplets $(s_1, s_2, y)$, where $s_1$ and $s_2$ are word sequences, and $y \in [0, 1]$ indicates the similarity between $s_1$ and $s_2$. Given sentence pair $s_1 = (w_1, w_2, \ldots, w_n)$ and $s_2 = (w_1, w_2, \ldots, w_m)$, where $n$ and $m$ are the length of $s_1$ and $s_2$, respectively. $s_1$ and $s_2$ are fed into the upper and lower branch network of our SNN and then the output of the SNN corresponding to $s_1$ and $s_2$ are described as:

\[
h'_1 = [h'_1; h'_2]
\]
\[
    h'_2 = [h'_2, h'_2]
\]

Then, \(h'_1\) and \(h'_2\) are used as \(X\) and \(Y\) of Section 3.3 to calculate the interactive self-attentive weight. Finally, the vector for calculating the similarity is obtained by equations 13 and 14. Then, the similarity is measured by the Manhattan distance formula (i.e., equation 17), and the evaluation score on the test set is computed by the Pearson function.

\[
    Dis(C_x, C_y) = \exp(-\|C_x - C_y\|_1),
\]

where \(C_x\) and \(C_y\) are the outputs of interactive self-attention, as shown in equations 13 and 14.

**IV. EXPERIMENTS AND SETTINGS**

**A. DATASETS**

In our experiments, three datasets measuring the similarity of biomedical texts are used to verify the effectiveness of the proposed model and the comparison experiments of the baseline, namely, DBMI released by the N2C2-Clinical semantic textual similarity task challenge (https://n2c2.dbmi.hms.harvard.edu) and CDD [36] (according to different sources, namely, the PubMed Central database and Conserved Domain Database, named CDD-ful and CDD-ref) constructed by Islamaj et al. The DBMI dataset consists of the training set containing 1,655 sentence pairs and the test set containing 412 sentence pairs. However, CDD-ful and CDD-ref contain 2,571 and 2,588 sentence pairs, respectively, given in the form of query-answer groups. To conduct our experiments, we randomly split 20% of the training set of DBMI as the validation set. Furthermore, the answers in each query-answer group of CDD-ful/ref are shuffled. Then, for 5x cross-validation, 20%, 20% and 60% of each query-answer group are chosen as part of the test set, validation set, and training set, respectively. The final divided dataset of CDD-ful/ref is available at https://github.com/dllzg2012/ISA-SNN.git.

**B. EVALUATION METRICS**

1) **PEARSON CORRELATION COEFFICIENT**

The accuracy of the predicted results is evaluated by this method and reflects the degree of linear correlation between two variables [37]. Given \(P = (p_1, p_2, \ldots , p_n)\) and \(Y = (y_1, y_2, \ldots , y_n)\), where \(P\) is the set of scores for \(n\) sentence pairs generated by our model and \(Y\) is the set of scores of the human assessors’ judgment. Each \(y_i\) in the set \(Y\), represents the semantic textual similarity between the \(i^{th}\) sentence pair \(p_i\) in the set \(P\). The Pearson correlation coefficient \(r\) is defined as follows:

\[
    r = \frac{\sum_{i=1}^{n} (p_i - \bar{p})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (p_i - \bar{p})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}
\]

2) **SPEARMAN CORRELATION COEFFICIENT**

When the two set joint \((P, Y)\) is bivariate normal, the Pearson correlation coefficient provides a complete summary of the association between \(P\) and \(Y\). However, the Pearson correlation coefficient is extremely sensitive to even the slightest departures from normality—a single outlier can easily conceal the underlying association [38]. Therefore, the robust correlation coefficient, namely, the Spearman coefficient correlation, is used to estimate the similarity of sentence pairs. The Spearman correlation coefficient \(\rho\) is just a Pearson correlation coefficient \(r\) between the ranked variables, which is defined as follows.

\[
    \rho = \frac{\sum_{i=1}^{n} (r[p_i] - \bar{r}[p])(r[y_i] - \bar{r}[y])}{\sqrt{\sum_{i=1}^{n} (r[p_i] - \bar{r}[p])^2 \sum_{i=1}^{n} (r[y_i] - \bar{r}[y])^2}}
\]

The degree of correlation is divided into five levels, i.e., perfect (0.8-1.0), strong (0.6-0.8), moderated (0.4-0.6), weak (0.2-0.4), zero (0.0-0.2).

In addition, because the mean square error is adopted as the loss function in the training process, the mean square error (MSE) is also used as our evaluation criteria.

3) **PRECISION, RECALL, F1-SCORE, AND ACCURACY**

Measuring the similarity between sentences is also converted into a binary classification task for verifying generalization performance of the proposed method and other existing methods. To perform the experiment, both the annotated similarity score and prediction score are converted to a classification label. First, annotated scores of DBMI([0-5]) and CDD-ful/ref([1-5]) are mapped to a [0-1] interval via equation 20.

\[
    Map(x, a, b, c, d) = a + \frac{x - c}{d - c} \times (b - a)
\]

where \(a\) and \(b\) are the upper and lower boundaries, respectively, of the target interval. Additionally, \(c\) and \(d\) are the upper and lower boundaries, respectively, of the original interval. Here, \(a = 0\) and \(b = 1\), for the DBMI and CDD-ful/ref dataset, \(c = 0, d = 5\) and \(c = 1, d = 5\), respectively. Then, as in [35], the predicted label is 1 when the output \(Dis(C_x, C_y) \geq 0.5\). Otherwise, the predicted label is 0. Finally, to evaluate the performance of the binary classification task, accuracy, precision, recall, and the F1-score are adopted, which are defined as follows.

\[
    \text{Acc} = \frac{TP + TN}{TP + FP + TN + FN}
\]

\[
    P = \frac{TP}{TP + FN}
\]

\[
    R = \frac{TP}{TP + FP}
\]

\[
    F1 - \text{Score} = \frac{2 \times P \times R}{P + R},
\]

where \(TP, TN, FP\) and \(FN\) are the number of true positives, true negatives, false positives and false negatives, respectively.

**C. TRAINING DETAILS**

Our implementation of all models is based on the open source deep learning package Keras running on
TABLE 1. Performance comparison of our method and other existing methods.

| Method        | CDD-ref/ful | DBMI     |
|---------------|-------------|----------|
|               | Pearson     | Spearman | MSE  | Pearson     | Spearman | MSE  |
| MaLSTM        | 0.531/0.611 | 0.545/0.615 | 1.161/1.450 | 0.458 | 0.466 | 3.296 |
| SCNN          | 0.494/0.560 | 0.512/0.586 | 1.786/2.992 | 0.461 | 0.479 | 7.869 |
| ImprovedSNN   | 0.622/0.682 | 0.649/0.686 | 0.895/1.154 | 0.573 | 0.554 | 2.153 |
| AttentiveSNN  | 0.625/0.669 | 0.638/0.679 | 0.868/1.184 | 0.606 | 0.584 | 1.953 |
| IA-SNN        | 0.651/0.653 | 0.654/0.65 | 0.873/1.255 | 0.612 | 0.567 | 1.734 |
| ISA-SNN       | 0.658/0.713 | 0.664/0.719 | 0.803/1.057 | 0.656 | 0.614 | 1.512 |

V. RESULTS AND ANALYSIS
A. BASELINES AND OUR MODELS
To demonstrate the effectiveness of our 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 [10] and achieved state-of-the-art results on the SICK [39] corpus.
- **SCNN**: like MaLSTM, replaced the recurrent neural network with a convolutional neural network.
- **ImprovedSNN**: proposed by Chi and Zhang [35], it employed hierarchical attention [29] to give different words different attention weights and achieved better results on a large dataset downloaded from Stanford Web.
- **AttentiveSNN**: proposed by Bao et al. [13], it regarded the attention weights as the coefficient of the Manhattan distance and achieved a higher Pearson correlation score than other methods on a cross-lingual textual similarity corpus.
- **SNN**: our baseline, similar to MaLSTM, but double biLSTMs in each branch network are employed in the model.
- **IA-SNN**: our baseline introducing interactive attention into a Siamese neural network.
- **ISA-SNN**: the proposed model with integrated interactive self-attention (ISA) with a Siamese neural network.

B. PERFORMANCE COMPARISON WITH OTHER EXISTING METHODS
To show the validity of our model, we report results on the CDD-ref and DBMI corpus. The results obtained from applying the proposed model to the test sets are shown in Table 1. First, MaLSTM without the attention mechanism achieves the worst results. This indicates that the weights of words within a sentence obtained by the attention mechanism are useful for distinguishing textual similarity. Second, compared with ImprovedSNN, the Pearson, Spearman correlation coefficient gained by the AttentiveSNN increase by 0.03 on DBMI. However, the Spearman correlation coefficient decreases by 0.01, while the Pearson correlation coefficient only increases by 0.003 on CDD-ref. This shows that the attention mechanism with a randomly initialized context vector introduces noise, leading to a decline in discrimination ability. Third, the performance of IA-SNN outperforms the other four methods, which reveals the importance of interactive semantic information for estimating semantic similarity between sentences. However, self-attention helps improve the performance of ISA-SNN, attaining a 0.656 Pearson correlation coefficient on DBMI. Finally, the scope of the performance gains on CDD-ref and DBMI shows the difficulty of the CDD-ref dataset. Choosing an appropriate model for this dataset is necessary.

For a more detailed analysis of the proposed attention’s ability in assessment similarity between biomedical sentences, the Manhattan distance of the worst and best sentence pairs is shown in Table 2 (enlarged to a similarity scale 1–5 via equation 20). The first sentence pair, containing 93 tokens, is the worst dissimilar example. The IA-SNN obtains the best score 2.7 (predicted score 0.45), and ISA-SNN achieves 3.4. This example shows that interactive self-attention may cause slight noise for long sentences. However, the second example, consisting of one sentence with 42 tokens and another with 14 tokens, is the best dissimilar sentence pair. The Manhattan distance obtained by ISA-SNN is obviously better than other methods including IA-SNN. Finally, both the third and fourth instances are the most similar sentence pairs. Moreover, the
query is exactly the same as the answer sentence, and all methods achieve a similar score on the two sentence pairs (i.e., the third pair is the worst and the fourth is best). The scores of the two sentence pairs show that IA-SNN and ISA-SNN perform on the most similar sentence pairs better than other methods owing to the enhanced sentence semantic representation via interactive attention.

We implement the hybrid system [40] without additional features, which is used for the OHNLP 2018 task2, to further verify the effectiveness of ISA. First, the attention mechanism in ABCNN [41] is replaced with ISA, named ISA-BCNN. ABCNN achieves a MAP and MRR of 0.647 and 0.658, respectively, and the MAP and MRR of ISA-BCNN are 0.676 and 0.689, respectively. Second, ISA-BCNN serves as a substitute for ABCNN in [40]. Finally, the results of [40] and our implemented hybrid system are illustrated in Table 3. The performance of ISA-BCNN is better than that of ABCNN, which shows the effectiveness of ISA.

### C. THE EFFECT OF SELF-ATTENTION AND INTERACTIVE SELF-ATTENTION

We also investigate the effect of self-attention and interactive self-attention on the performance in our model, as shown in Table 4. First, the proposed method ISA-SNN achieves the best performance on CDD-ref and DBMI across all three evaluation metrics. In contrast, SA-SNN attains the best MSE score on the CDD-ful dataset. Second, self-attention results in MSE decrease by nearly 1.232 compared to the baseline, but interactive self-attention leads to MSE increasing by 0.002 relative to self-attention. The results indicate that the semantic features (attention weights) learned by self-attention contribute to promoting the semantic representation in a single sentence, but interactive attention introduces a small amount of noise on CDD-ful. Nevertheless, interactive self-attention yields the best results compared to self-attention on CDD-ref and DBMI. The main reason may be that interactive attention enlarges the semantic difference

### TABLE 2. Most similar/dissimilar sentence pairs.

| No | Sentence | G1 | G2 | G3 | G4 | G5 | G |
|----|----------|----|----|----|----|----|----|
| 1  | (Q) Moreover, UBA-like domain is required for binding ubiquitylated-protein substrates, UIM motif is responsible for the binding to cullin RING ligases (CRLs), and UBX domain is essential for p97 binding (32 tokens) | 3.5 | 3.6 | 3.6 | 2.7 | 3.4 | 1 |
| 2  | (Q) Periplasmic xylose-binding component of the ABC-type transport systems that belong to a family of pentose/hexose sugar-binding proteins of the type 1 periplasmic binding protein (PBP1) superfamily, which consists of two alpha/beta globular domains connected by a three-stranded hinge. (42 tokens) | 1.9 | 1.7 | 2 | 1.9 | 1.5 | 1.5 |
| 3  | (Q) Kalirin and Trio are encoded by separate genes in mammals and by a single one in invertebrates. (18 tokens) | 1.5 | 1.4 | 1.1 | 1.5 | 1.6 | 5 |
| 4  | (Q) Strong expression of the human gene and its mouse ortholog Acyl1 in brain, liver, and kidney suggest a role of the enzyme in amino acid metabolism of these organs. (32 tokens) | 4.1 | 4.5 | 4.1 | 5 | 4.8 | 5 |

*G1, G2, G3, G4, G5 denote predicted scores of MalSTM, ImprovedSNN, AttentiveSNN, IA-SNN, and ISA-SNN, respectively and G denotes the manual score.

### TABLE 3. Performance comparison of ISA-BCNN and ABCNN on DBMI.

| Method | Pearson |
|--------|---------|
| ABCNN1-ML | 0.557 |
| ABCNN2-ML | 0.554 |
| ABCNN3-ML | 0.562 |
| ABCNN1-DL | 0.487 |
| ABCNN2-DL | 0.566 |
| ABCNN3-DL | 0.541 |
| ISA-BCNN1-ML | 0.571 |
| ISA-BCNN2-ML | 0.566 |
| ISA-BCNN3-ML | 0.575 |
| ISA-BCNN1-DL | 0.501 |
| ISA-BCNN2-DL | 0.592 |
| ISA-BCNN3-DL | 0.619 |

*ML denotes using machine learning regression for prediction, and DL denotes using the fully connected layer of the network for prediction. ABCNN contains attention operation in Convolution input layer (named ABCNN1) and pooling layer (named ABCNN 2). ABCNN3 combines ABCNN1 and ABCNN2 by stacking them.*
TABLE 4. The effect of self-attention and interactive self-attention on CDD-ful/-ref and DBMI.

| Method     | CDD-ful/-ref       | DBMI       |
|------------|--------------------|------------|
|            | Pearson  | Spearman  | MSE       | Pearson  | Spearman  | MSE       |
| SNN        | 0.511/0.531 | 0.518/0.545 | 2.287/1.161 | 0.458 | 0.466 | 3.296 |
| SA-SNN     | **0.717/0.622** | **0.722/0.649** | **1.055/0.894** | 0.573 | 0.554 | 2.153 |
| ISA-SNN    | 0.713/0.658 | 0.719/0.664 | 1.057/0.803 | **0.656** | **0.614** | **1.512** |

**FIGURE 2.** Visualization of the attention score between ISA-SNN and SA-SNN. (a) Similar sentence pair (between “Patient has contact information and understands the need to call with any questions or concerns.” and “The patient knows how to reach us any time with questions or concerns.”). ISA-SNN and SA-SNN achieve a Manhattan distance of 1 and 0.92, respectively; (b) dissimilar sentence pair (between “The GoLYTELY was then dispersed to the research volunteer for use in this study.” and “Patient was brought to the EP suite in the fasting state.”). ISA-SNN and SA-SNN achieve a Manhattan distance of 0.04 and 0.12, respectively.

between dissimilar sentences and reduces the semantic difference between similar sentences, which results in improving the discrimination ability in textual similarity.

### D. ANALYSIS OF ATTENTION SCORE VISUALIZATION BETWEEN ISA-SNN AND SA-SNN

To compare the difference between self-attention and interactive self-attention, we still visualize the attention score of similar and dissimilar sentence pairs, as shown in Figure 2, Tables 5 and 6. Figure 2(a) shows that comparisons of attention score on a similar sentence pair, namely, “Patient has contact information and understands the need to call with any questions or concerns.” (sentence A) and “The patient knows how to reach us any time with questions or concerns.” (sentence B). For the ISA-SNN, the words of maximum and minimum attention weight in sentence
TABLE 5. Maximum, minimum attention score, difference value of single sentence and absolute difference value of sentence A’s difference value and B’s difference value in similar pair.

| Method | Sentence Pair | max.          | min.          | diff.         | \(|\text{diff}_A - \text{diff}_B|\) |
|--------|---------------|---------------|---------------|---------------|----------------------------------|
|        | A             | 0.00499354    | 0.00497812    | 0.00001542    |                                  |
|        | B             | 0.00499167    | 0.00497642    | 0.00001525    | 0.00000017                      |
| ISA-SNN| A             | 0.00499238    | 0.00497568    | 0.00001670    |                                  |
|        | B             | 0.00499035    | 0.00497615    | 0.00001420    | 0.00000025                      |
|        | A             | 0.00516566    | 0.00500083    | 0.00016483    |                                  |
|        | B             | 0.00500344    | 0.00500027    | 0.00000017    | 0.00016166                      |
| SA-SNN | A             | 0.00508758    | 0.00505297    | 0.00003461    |                                  |
|        | B             | 0.00502431    | 0.00500124    | 0.00002307    | 0.00001154                      |

* Similar sentence pair 1 (between “Patient has contact information and understands the need to call with any questions or concerns.” and “The patient knows how to reach us any time with questions or concerns.”).

* Similar sentence pair 2 (between “Heart: S1/S2 regular rate and rhythm, without murmurs, gallops or rubs.” and “Heart: S1, S2 regular rate and rhythm, no abnormal heart sounds or murmurs.”).

max, min denote the maximum and minimum attention score of one sentence, respectively. diff. denotes the difference value of maximum and minimum attention score. \(|\text{diff}_A - \text{diff}_B|\) denotes absolute difference value of sentence A’s diff. and sentence B’s diff.

TABLE 6. Maximum, minimum attention score, difference value of single sentence and absolute difference value of sentence A’s difference value and B’s difference value in dissimilar pair.

| Method | Sentence Pair | max.          | min.          | diff.         | \(|\text{diff}_A - \text{diff}_B|\) |
|--------|---------------|---------------|---------------|---------------|----------------------------------|
|        | A             | 0.00500326    | 0.00498583    | 0.00001743    |                                  |
|        | B             | 0.00506662    | 0.0049866     | 0.00007802    | 0.00006059                      |
| ISA-SNN| A             | 0.00503286    | 0.005001     | 0.00003186    |                                  |
|        | B             | 0.00514404    | 0.00501409    | 0.00012995    | 0.00009809                      |
|        | A             | 0.00509521    | 0.0050426     | 0.00005261    |                                  |
|        | B             | 0.00511322    | 0.00505056    | 0.00006266    | 0.00010005                      |
|        | A             | 0.00502815    | 0.00505063    | 0.00002252    |                                  |
|        | B             | 0.00515039    | 0.00506878    | 0.00008161    | 0.00009599                      |

* Dissimilar sentence pair 1 (between “The GoLYTELY was then dispersed to the research volunteer for use in this study.” and “Patient was brought to the EP suite in the fasting state.”).

* Dissimilar sentence pair 2 (between “Rotation was set according to a combination of the transepicondylar axis and the AP axis of the femur,” and “Palpation was also performed to note the location of the occipital artery.”).

A are “with” (attention score: 0.00499354) and “Patient” (attention score: 0.00497812), respectively. Additionally, the maximum and minimum attention weights in sentence B are 0.00499167 and 0.00497642. Thus, the difference values of sentences A and B are 0.00001542 and 0.00001525. However, the attention weights of SA-SNN ranged from 0.00500083 to 0.00516566 in sentence A and from 0.00500124 to 0.00500124 in sentence B.

Therefore, the difference values of sentences A and B are 0.00016483 and 0.00001525, respectively. Due to the above difference, the Manhattan distance of 0.92 and 1 are obtained by SA-SNN and ISA-SNN, respectively. The SA-SNN and ISA-SNN both obtain the conclusion that the two sentences are completely or mostly equivalent (level is perfect). Additionally, attention visualization of dissimilar sentence pairs, namely, “The GoLYTELY was then dispersed to the research volunteer for use in this study.” (sentence A) and “Patient was brought to the EP suite in the fasting state.” (sentence B), is shown in Figure 2(b). The maximum and minimum attention weights obtained by ISA-SNN are 0.00500326 and 0.00498583 in sentence A. Moreover, the range of attention weights is from 0.00500083 to 0.00516566 in sentence B. The difference values of sentences A and B are 0.00001743 and 0.00007802, respectively. In contrast, the SA-SNN achieved difference values of 0.00005261 and 0.00006266 according to the range of attention weights, which fluctuate between 0.0050426 and 0.00509521 in sentence A, and between 0.00505056 and 0.00511322 in sentence B separately. For similar sentence pair, absolute difference value between sentence A’s difference value and sentence B’s difference value obtained by ISA-SNN is smaller than that obtained by SA-SNN. However, for dissimilar sentence pair, the...
TABLE 7. Performance comparison of our method and baseline when regarding as binary classification task.

| Method   | CDD-ful/-ref | DBMI       |
|----------|--------------|------------|
|          | P           | R          | F1-score | accuracy | P       | R       | F1-score | accuracy |
| SNN      | 0.81/0.654  | 0.33/0.744 | 0.469/0.696 | 0.629/0.695 | 0.3957  | 0.875  | 0.5450   | 0.5450   |
| SA-SNN   | 0.798/0.719 | 0.804/0.666 | 0.801/0.692 | 0.802/0.721 | 0.4864  | 0.8437 | 0.6171   | 0.6739   |
| ISA-SNN  | 0.792/0.728 | 0.747/0.709 | 0.769/0.718 | 0.777/0.739 | 0.6273  | 0.7890 | 0.6989   | 0.7883   |

opposite conclusion is obtained, as shown in Tables 5 and 6. These results reveal that the difference between the maximum and minimum values is reduced in similar pairs but is enlarged in the dissimilar sentence pairs. Therefore, interactive self-attention is useful to reduce the semantic difference of the similar sentence pair and enlarge the semantic difference of the dissimilar sentence pair.

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

Estimating the similarity between sentences is also regarded as the sentence pair classification task. The performance of the converted binary classification task, as mentioned above, is illustrated in Table 7. The SNN achieves the best recall rate, but the proposed method shows the best F1-score and accuracy on DBMI and CDD-ref. However, SA-SNN obtains the best recall rate, F1-score, and accuracy on CDD-ful, and ISA-SNN obtains the best precision and recall rate on CDD-ref and DBMI. The SNN extracts the semantic information between sentences via shared weights, so there is almost no noise. In contrast, the introduction of self-attention and interactive attention reduces the number of false negative examples but causes a small amount of noise, which increases the number of false positive examples, resulting in a decrease in recall rate. In terms of results, despite a small decrease in the recall rate, compared to the baseline, the F1-scores and accuracy have been significantly improved, with 0.022 and 0.044 improvement on the CDD-ref corpus, and 0.153 and 0.243 improvement on the DBMI corpus. However, on CDD-ful, the performance of SA-SNN is better than that of ISA-SNN, the reasons are given later.

Moreover, to analyze the generalization performance of three methods, we draw the ROC curve and compute the AUC score on the DBMI dataset, as shown in Figure 3. It demonstrates that ISA-SNN outperforms the other methods in the classification task and is more suitable for applying to other sentence pair datasets.

F. ANALYSIS OF CDD-FUL/-REF

The aforementioned results show that the similarity and classification performance of SA-SNN on CDD-ful outperforms that of ISA-SNN. The direct reason may be that the interaction semantic information causes a slight amount of noise in this corpus. However, on CDD-ref corpus found in the same literature [36], ISA-SNN performs better. Therefore, we compare the difference between the two datasets according to the labeled score distribution, sentence length, and text quality.

First, we give statistics on the similarity levels distribution of the judged pairs of sentences in the CDD-ful/-ref. Figure 4 shows that the number of sentences with 4-5 scores in CDD-ful is almost twice that in CDD-ref while the number of sentences with 2-3 points is the opposite. Second, as shown in Figure 5, the number of sentence pairs with more than 120 words in CDD-ful is more than that in the CDD-ref, while 1-40 is the opposite. Finally, more special tokens such as “Figure”, “()” are found in CDD-ful. Furthermore, part of the answer is the same as the query among query-answer
groups such as query numbers 1983390001 and 1935450007. These differences may be due to the different sources of CDD-ful and CDD-ref (i.e., PubMed Central database and Conserved Domain Database according to the literature [35]). Hence, the reason resulting in the different performance between the two datasets may be that i) interaction semantic information causes noise when it handles similar or long sentences, which results in the reduction in discrimination ability between sentences; and ii) interactive attention slightly reduces the ability of estimation similarity on corpora with noise.

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose an interactive self-attention mechanism and introduce it into the Siamese neural network with bidirectional LSTM units (named ISA-SNN), which can enhance the semantic expression in a single sentence via self-attention on the basis of the shared parameters of the Siamese network. Furthermore, inherent interactive semantic information between sentences is learned by means of interactive attention, which contributes to reducing the semantic difference in a similar sentence pair and enlarging the semantic difference in a dissimilar sentence pair. Moreover, we conduct experiments on three biomedical datasets. Experimental results indicate that the proposed model for measuring biomedical textual similarity has higher performance on the three datasets. However, further improvement is needed in long sentences and noisy datasets. Therefore, an investigation into whether the multi-head self-attention mechanism may optimize our model will be considered in future work. In addition, external biomedical knowledge may be combined 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, 2013, Art. no. e65814.
[2] M. Ben Aouicha and M. A. H. Taieb, “Computing semantic similarity between biomedical concepts using new information content approach,” J. Biomed. Inform., vol. 59, pp. 258–275, Feb. 2016.
[3] R. Mihalcea, C. Corley, and C. Strapparava, “Corpus-based and knowledge-based measures of text semantic similarity,” in Proc. AAM, 2005, pp. 775–780.
[4] G. Glavaš, F. Šaric, M. Karan, and J. Šnajder, “Tagalog: Systems for measuring semantic text similarity,” in Proc. 1st Joint Conf. Lexical Comput. Semantics, 2012, pp. 441–448.
[5] T. Kenter and M. de Rijke, “Short text similarity with word embeddings,” in Proc. 24th ACM Int. Conf. Inf. Knowl. Manage. (CIKM), vols. 19–23, 2015, pp. 1411–1420.
[6] D. Merks and S. Frank, “Learning semantic sentence representations from visually grounded language without lexical knowledge,” 2019, arXiv:1903.11393. [Online]. Available: http://arxiv.org/abs/1903.11393
[7] 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.
[8] 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: http://arxiv.org/abs/1810.08740
[9] P. Neculoiu, M. Verseeth, and M. Rotaru, “Learning text similarity with Siamese recurrent networks,” in Proc. 1st Workshop Represent. Learn. (NLP), Jan. 2016, pp. 148–157.
[10] J. Mueller, “Siamese recurrent architectures for learning sentence similarity,” in Proc. AAAI, Nov. 2018, pp. 2786–2792.
[11] 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, 2018, Art. no. e0193919.
[12] E. L. Pontes, S. Huet, A. C. Linhares, and J.-M. Torres-Moreno, “Predicting the semantic textual similarity with Siamese CNN and LSTM,” 2018, arXiv:1810.10641. [Online]. Available: http://arxiv.org/abs/1810.10641
[13] 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.
[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017, arXiv:1706.03762. [Online]. Available: http://arxiv.org/abs/1706.03762
[15] D. Ma, S. Li, X. Zhang, and H. Wang, “Interactive attention networks for aspect-level sentiment classification,” 2017, arXiv:1709.00893. [Online]. Available: http://arxiv.org/abs/1709.00893
[16] D. Hiemstra, “Term-specific smoothing for the language modeling approach to information retrieval: The importance of a query term,” in Proc. 25th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR), vol. 2, 2002, pp. 35–41.
[17] H. Moen, F. Ginter, E. Marsi, L.-M. Peltonen, T. Salakoski, and S. Salanterä, “Care episode retrieval: Distributional semantic models for information retrieval in the clinical domain,” BMC Med. Inform. Decis. Making, vol. 15, no. S2, pp. 1–19, Dec. 2015.
[18] K. Papineni, S. Roukos, T. Ward, and W. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics, Jul. 2002, pp. 311–318.
[19] C. Corley and R. Mihalcea, “Measuring the semantic similarity of texts,” in Proc. ACL Workshop Empirical Model. Semantic Equivalence Entailment (EMSEE), 2008, pp. 13–18.
[20] W. Deng, H. Liu, J. Xu, H. Zhao, and Y. Song, “An improved quantum-inspired differential evolution algorithm for deep belief network,” IEEE Trans. Instrum. Meas., early access, Mar. 25, 2020, doi: 10.1109/TIM.2020.2983233.
[21] H. Zhao, J. Zheng, W. Deng, and Y. Song, “Semi-supervised broad learning system based on manifold regularization and broad network,” IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 67, no. 3, pp. 983–994, Mar. 2020.
[22] S. Chaudhuri, G. Polatkan, R. Ramanath, and V. Mithal, “An attentive survey of attention models,” 2019, arXiv:1904.02874. [Online]. Available: http://arxiv.org/abs/1904.02874
[23] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014, arXiv:1409.0473. [Online]. Available: http://arxiv.org/abs/1409.0473
[24] T. Zhang and M. Huang, “Learning structured representation for text classification via reinforcement learning,” in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 6053–6060.
[25] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, and J. Wang, “An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition,” Bioinformatics, vol. 34, no. 8, pp. 1381–1388, Apr. 2018.
[26] Y. Song, J. Wang, T. Jiang, Z. Liu, and Y. Rao, “Attentional encoder network for targeted sentiment classification,” 2019, arXiv:1902.09314. [Online]. Available: http://arxiv.org/abs/1902.09314
ZHENGGUANG LI received the master’s degree in computer science from Dalian Jiaotong University, in 2007. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Dalian University of Technology. His research interests include text mining and natural language processing in social media and biomedical fields.

HONGFEI LIN received the B.Sc. degree from Northeast Normal University, in 1983, the M.Sc. degree from the Dalian University of Technology, in 1992, and the Ph.D. degree from Northeastern University, in 2000. He is currently a Professor with the School of Computer Science and Technology, Dalian University of Technology. He is also the Director of the Information Retrieval Laboratory, Dalian University of Technology. He has published over 100 research articles in various journals, conferences, and books. His research interests include information retrieval, text mining, natural language processing, and effective computing. In recent years, he has focused on text mining for biomedical literature, biomedical hypothesis generation, information extraction from large biomedical resources, learning to rank, sentiment analysis, and opinion mining. His research projects are funded by the National Natural Science Foundation of China and the National High-Tech Development Plan.

WEI ZHENG received the Ph.D. degree in computer application technology from the Dalian University of Technology, China, in 2018. She is currently a Lecturer with the School of Software Engineering, Dalian Jiaotong University, Dalian, China. Her research interests include text mining for biomedical literature and social media, knowledge graph construction, and natural language processing.

MICHAEL M. TADESSE received the B.S. degree in management information systems (MIS) from Unity University College, Addis Ababa, Ethiopia, in 2005, and the M.S. degree in software engineering from Chongqing University, in 2011. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Dalian University of Technology. His current research interests include personality prediction, recommendation system information retrieval, and natural language processing.

ZHIHAO YANG received the Ph.D. degree in computer science from the Dalian University of Technology, China, in 2008. He is currently a Professor with the School of Computer Science and Technology, Dalian University of Technology. He has published more than 20 research articles on topics in biomedical literature data mining. His current research interests include biomedical literature data mining and information retrieval.

JIAN WANG received the B.S., M.S., and Ph.D. degrees from the Dalian University of Technology, China, in 1988, 1993, and 2014, respectively. She is currently a Professor with the School of Computer Science and Technology, Dalian University of Technology. Her current research interests include biomedical literature data mining, information retrieval, and natural language processing.