IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 31, 2019, accepted November 12, 2019, date of publication November 21, 2019,
date of current version December 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2954988

Implicit Discourse Relation Recognition via a BiLSTM-CNN Architecture With Dynamic Chunk-Based Max Pooling

FENGYU GUO[1,2], RUIFANG HE[1,2], AND JIANWU DANG[1,2,3], (Member, IEEE)
[1]College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[2]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin 300350, China
[3]School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa 9231292, Japan

Corresponding author: Ruifang He (rfhe@tju.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61976154, in part by the Tianjin Natural Science Foundation under Grant 18JCYBJC15500, in part by the National Key Research and Development Program of China under Grant 2018YFB1305200, and in part by the Tianjin Municipal Science and Technology Project under Grant 18ZXZNGX00330.

ABSTRACT Implicit discourse relation recognition is a serious challenge in discourse analysis, which aims to understand and annotate the latent relations between two discourse arguments, such as temporal and comparison. Most neural network-based models encode linguistic features (such as syntactic parsing and position information) as embedding vectors, which are prone to error propagation due to unsuitable pre-processing. Other methods apply different attention or memory mechanisms, mainly considering the key points in the discourse, yet ignore some valuable clues. In particular, those using convolution neural networks retain local contexts but lose word order information due to the standard pooling operation. The methods that use bidirectional long short-term memory network consider the word sequence and retain the global information, but cannot capture the context with different range sizes. In this paper, we propose a novel **D**ynamic **C**hunk-based Max Pooling **B**iLSTM-**CNN** framework (**DC-BCNN**) to address these issues. First, we exploit BiLSTMs to capture the semantic representations of discourse arguments. Second, we adopt the proposed convolutional layer to automatically extract the ''multi-granularity'' features (just like n-gram) by setting different convolution filter sizes. Then, we design a dynamic chunk-based max pooling strategy to obtain the important scaled features of different parts in one discourse argument. This strategy can dynamically divide each argument into several segments (called chunks) according to the argument length and the number of current pooling layer in the CNN and then select the maximum value of each chunk to indicate crucial information. We further utilize a fully connected layer with a softmax function to recognize discourse relations. The experimental results on two corpora (i.e., PDTB and HIT-CDTB) show that our proposed model is effective in implicit discourse relation recognition.

INDEX TERMS Implicit discourse relation recognition, discourse argument representation, dynamic chunk-based max pooling, BiLSTM, CNN.

I. INTRODUCTION

Discourse relation describes how two adjacent text units (e.g., clauses, sentences, and larger sentence groups) are connected logically, which can capture essential structural and semantic aspects of a discourse. A discourse relation instance is usually defined as a connective taking two arguments (as *Arg*1 and *Arg*2, respectively). Discourse relation recognition is significant for understanding discourse and beneficial to many downstream natural language processing (NLP) applications, e.g., machine translation [1], text summarization [2], as well as conversation systems [3], [4].

The task of automatically identifying discourse relations is relatively simple when explicit connectives such as *but* and *because* are given. Recognizing the implicit relations has been shown to be the performance bottleneck of discourse parser due to the lack of explicit connectives. Implicit discourse relations outnumber explicit relations in naturally occurring text; more than 80% of the words in Chinese discourse and 52% in English are implicit [5].

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

Therefore, this work concentrates on implicit discourse relation recognition that needs to infer the discourse relations from the semantic understanding of the specific contexts.

The concerned task can be straightforwardly formalized as a sentence-pair classification problem, which needs to infer the relations based on the two arguments. There are two questions that arise: how can discourse arguments be modelled properly and how can the interactions between arguments be captured.

For these issues, considerable researches have been performed for implicit discourse relation recognition with the use of traditional NLP linguistically informed features and machine learning algorithms [6]–[9]. They might be subject to the error propagation introduced by the imperfect quality of the supervised NLP toolkits. However, implicit discourse relations are rooted in semantics, which may make them hard to recover from surface features; thus, those feature-based methods did not report satisfactory performance. Recently, various neural network-based models have shown competitive results on this task, including convolutional neural networks (CNNs) [10], [11], and recurrent neural networks (RNNs) [13]–[15]. More researches based on basic neural networks exploit attention mechanisms, memory mechanisms and gate mechanisms to capture more complicated information from discourse arguments [11], [16]–[19]. Although these studies have improved performance to some extent, they are too complex to reproduce. A study [20] has shown that simple networks might outperform tree LSTM-based models. Meanwhile, many studies of Chinese discourse [21], [22] focus on macro level discourse structure analysis and utilize various kinds of features sets to improve the performance of local models, ignoring the generalization of models. They also claim that the training data for implicit discourse relation classification is too small to run powerful neural networks, which motivates us to address this task by a simple and effective method.

We argue that implicit discourse relation recognition needs to learn contextual information, which can obtain the semantic understanding of discourse. For example, suppose that we want to classify the discourse relation of the following pair (referred to as *Arg*1 in *italics* and *Arg*2 in **bold** throughout this paper):

*Arg*1: *You are really lucky*.

*Arg*2: **The earthquake came five hours after you left**.

If only taking the (lucky, earthquake) pair, it might construe a Comparison relation due to the contrasting sentiment polarity of the word pair. It is understood as a Contingency relation when the entire context of both arguments is considered. RNN can better capture the contextual information of the text, and we adopt BiLSTM as a variant of RNN to encode the semantic representation of discourse arguments.

In addition, the semantic representations of discourse argument with different granularities have different semantic understandings. For example, in the sentence "a sunset stroll along the South Bank affords an array of stunning vantage points." When we analyse the word "Bank", we may not know whether it means a financial institution or the land beside a river. The phrase "South Bank" may mislead us to take it as a financial institution. After obtaining the greater context "stroll along the South Bank", we can easily understand its real meaning. Convolutional neural networks that can recognize specific classes of n-grams and induce more abstract representations are a natural combination, which could obtain more effective representations. We can incorporate various window sizes for convolutional filters, allowing the network to capture wider ranges of n-grams to help with implicit discourse relation classification.

The convolutional neural networks typically utilize a standard max pooling layer that applies a max operation over a feature map to capture the most useful information. This operation may lose valuable facts such as word order information which is help identify implicit discourse relation. To reserve more effective information, some studies [12], [13], [23], [25] design improved pooling operations, such as dynamic pooling, dynamic k-max pooling,, as well as dynamic multi-pooling, to take more maximum values to retain information, but the word order information is still not effectively retained.

*S*1: 早上下了一场小雨，**校园里的小路都湿了。**

There was a gentle rain this morning,

The roads on the campus were wet.

*S*2: 校园里的小路都湿了，**早上下了一场小雨。**

The roads on the campus were wet,

There was a gentle rain this morning.

However, the word order clues of discourse has a considerable influence on relation understanding, especially in Chinese. Sentences *S*1 and *S*2 have changed the order of arguments, and they have different relations, namely, *S*1 might be a Causal relation and *S*2 construes an Explanation relation. We devise a dynamic chunk-based max pooling to select the maximum value of each part in the discourse argument, which preserves the relevant word order of the argument.

Success in speech recognition [54], [55], biomedical engineering [52], [53] and text tasks [40], [51] shows that the combination of RNN and CNN could obtain more comprehensive clues. In our task, both the contextual information of discourse and the local cues of different discourse units are significant to identify implicit discourse relation.

Therefore, we address implicit discourse relation recognition for both English and Chinese, and propose a novel **D**ynamic **C**hunk-based Max Pooling **Bi**LSTM-**CNN** (i.e., DC-BCNN) framework. Specifically, we encode two discourse arguments by bidirectional long short-term memory networks to reserve the contextual information, and then adopt convolutional neural network to extract the semantic features of different n-grams by different sizes of filters. In the CNN module, the conventional max pooling layer utilizes

a max operation over the obtained feature map to select the most useful information, which may miss the valid information in the other parts of one argument, for example, word order which may directly influence the relation identification. Inspired by previous work [24], [25], we design a dynamic chunk-based max pooling operation to divide the arguments into several segments according to the length of arguments and the structure of the neural network and then select the maximum value of each segment to retain as much information as possible without other manual operations. Finally, a classifier is trained to identify the discourse relations.[1]

In summary, our main contributions are as follows:

- We propose a novel DC-BCNN framework for implicit discourse relation recognition, which can automatically induce the semantic understanding from the wider ranges of n-grams and reserve more valid information without complicated NLP pre-processing;
- We design a dynamic chunk-based max pooling operation to capture more valuable information within two discourse arguments for the task;
- We conduct a series of experiments on English and Chinese corpora to evaluate the effectiveness of our proposed model.

The rest of this paper is organized as follows: Section II elaborates the main framework and the detailed description of dynamic chunk-based max pooling operation. The experimental preparation and the compared baselines are shown in Section III. Section IV provides the details of experimental results and discussion. Section V describes the related work towards implicit discourse relation recognition. The conclusion and future work are presented in Section VI.

II. THE PROPOSED APPROACH

In this section, we first give an overview of our DC-BCNN framework, as shown in Figure 1. Our framework primarily involves the following components: (i) embedding layer, which contains lexical information for each word and trained from an external corpus in an unsupervised manner; (ii) discourse argument representation, which can learn the historical and future abstractive representation by BiLSTMs; (iii) multi-granularity feature extraction, which proposes a dynamic chunk-based max pooling CNN to learn the compositional semantic features; (iv) a classifier output, which calculates the confidence scores for the discourse relations. Next, we will illustrate the details of our proposed model.

A. EMBEDDING LAYER

Unlike the high-dimensional and sparse of one-hot encoded vector, distributed representation is low-dimensional, learned continuous vector, which reflects the much more sophisticated relationships between words. The idea of distributed representation for symbolic data is one of the most important reasons why the neural network works. Distributed representation was proposed by Hinton and has been a popular research topic for more than twenty years [26], [27], [30]. Formally, the embedding layer can be a simple project layer where word embedding is achieved by lookup table operation according to the indexes.

To model two discourse arguments, we transform the one-hot representation of argument into a distributed representation. We associate each word w in the vocabulary with a vector representation $\boldsymbol{x}_w \in \mathbb{R}^d$, where d is the dimension of the embeddings. Since each argument is viewed as a sequence of word vectors, let $\boldsymbol{x}_i^1(\boldsymbol{x}_i^2)$ be the i-th word vector in $Arg1$ ($Arg2$); the arguments in a discourse relation are expressed as:

$$\begin{aligned} Arg1: &\ [\boldsymbol{x}_1^1, \boldsymbol{x}_2^1, \ldots, \boldsymbol{x}_{n_1}^1], \\ Arg2: &\ [\boldsymbol{x}_1^2, \boldsymbol{x}_2^2, \ldots, \boldsymbol{x}_{n_2}^2]. \end{aligned} \tag{1}$$

where $Arg1$ ($Arg2$) has n_1 (n_2) words.

B. DISCOURSE ARGUMENT REPRESENTATION

The Long Short-Term Memory network (LSTM) [28] is a variant of the recurrent neural network, and specifically addresses the issue of learning long-term dependencies. Considering that it is good at modelling a sequence of words with contextual information, we adopt it to encode two discourse arguments. Given the word representations of the arguments as we just described, the LSTM computes the state sequence for each position t using the following equations:

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i[\boldsymbol{x}_t, \boldsymbol{h}_{t-1}] + \boldsymbol{b}_i), \tag{2}$$
$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f[\boldsymbol{x}_t, \boldsymbol{h}_{t-1}] + \boldsymbol{b}_f), \tag{3}$$
$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o[\boldsymbol{x}_t, \boldsymbol{h}_{t-1}] + \boldsymbol{b}_o), \tag{4}$$
$$\tilde{\boldsymbol{c}_t} = \tanh(\boldsymbol{W}_c[\boldsymbol{x}_t, \boldsymbol{h}_{t-1}] + \boldsymbol{b}_c), \tag{5}$$
$$\boldsymbol{c}_t = \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}_t} + \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1}, \tag{6}$$
$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t). \tag{7}$$

where $\boldsymbol{i}_t, \boldsymbol{f}_t, \boldsymbol{o}_t, \boldsymbol{c}_t, \boldsymbol{h}_t$ denote the input gate, forget gate, output gate, memory cell and hidden state at position t, respectively. $\boldsymbol{W}_i, \boldsymbol{W}_f, \boldsymbol{W}_o, \boldsymbol{W}_c, \boldsymbol{b}_i, \boldsymbol{b}_f, \boldsymbol{b}_o, \boldsymbol{b}_c$ are the parameters of the neural network. Besides, [ ] means the concatenation operation of $\boldsymbol{x}_t, \boldsymbol{h}_{t-1}$ vectors, and σ denotes the logistic sigmoid function and $\odot$ denotes the element-wise product.

Notice that LSTM only considers the context from the past, but the contextual information from the future can also be crucial. Referring to the previous work, we implement a bidirectional LSTM (BiLSTM) neural network to model the argument sequences. BiLSTM preserves both the historical and future information by two separate LSTMs in the forward and reverse directions. Therefore, we can obtain two representations $\overrightarrow{\boldsymbol{h}_t}$ and $\overleftarrow{\boldsymbol{h}_t}$ at each position t of the sequence. Then, we concatenate them to obtain the intermediate state $\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}_t}, \overleftarrow{\boldsymbol{h}_t}]$. As shown in Figure 1, we encode $Arg1$ and $Arg2$ into the contextual representations by two Bi-LSTMs; namely, $\boldsymbol{h}_i^1 = [\overrightarrow{\boldsymbol{h}_i^1}, \overleftarrow{\boldsymbol{h}_i^1}]$ and $\boldsymbol{h}_j^2 = [\overrightarrow{\boldsymbol{h}_j^2}, \overleftarrow{\boldsymbol{h}_j^2}]$ are the intermediate states of the i-th word in $Arg1$ and the j-th word in $Arg2$, respectively, where $\overrightarrow{\boldsymbol{h}_i^1}, \overrightarrow{\boldsymbol{h}_j^2} \in \mathbb{R}^d$ and $\overleftarrow{\boldsymbol{h}_i^1}, \overleftarrow{\boldsymbol{h}_j^2} \in \mathbb{R}^d$ are the outputs from two directions.

[1] We'll provide our code on https://github.com/w6688j/ after organizing.

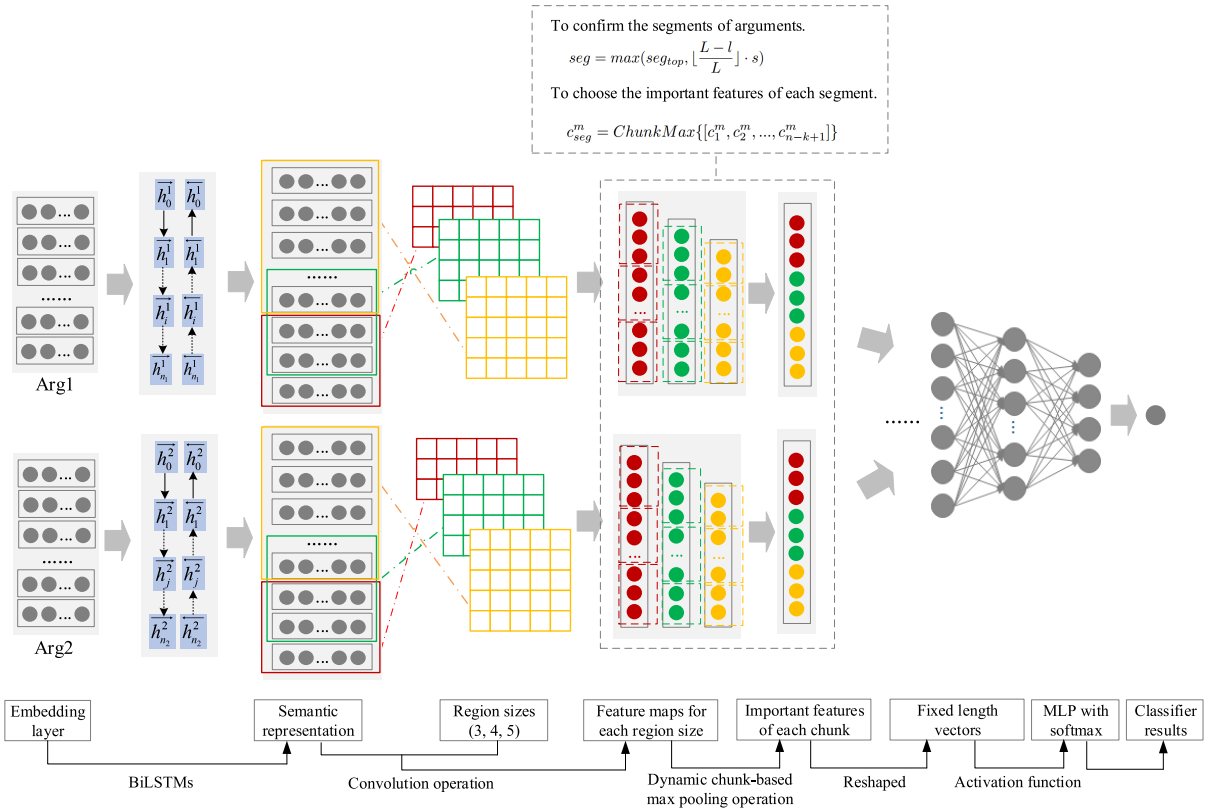


FIGURE 1. The dynamic chunk-based max pooling BiLSTM-CNN framework (DC-BCNN).

C. MULTI-GRANULARITY FEATURE EXTRACTION

After obtaining the argument representation with contextual information, we expect to capture the complicated and various features of two discourse arguments. Therefore, we devise a new variant of the standard CNN, a dynamic chunk-based max pooling CNN, to retain more semantics of the arguments.

1) CONVOLUTION LAYER

Compared to some previous models that apply a single-window size, we utilize a number of filters $W \in \mathbb{R}^{k \times d}$ that summarize the information of the k-word window and produce a new feature. For the window of k words $\boldsymbol{h}_{i:i+k-1}$, a filter $\boldsymbol{f}_m$ ($1 \leq m \leq M$, where M denotes the number of filters) is used to generate the feature $\boldsymbol{c}^m_i$ corresponding to a certain n-gram representation. The formula is defined as follows:

$$\boldsymbol{c}^m_i = \sigma(\boldsymbol{w} \cdot \boldsymbol{h}_{i:i+k-1} + \boldsymbol{b}). \tag{8}$$

where σ is a non-linear function (e.g., sigmoid, tanh and ReLU), $\boldsymbol{b} \in \mathbb{R}$ is a bias term, and $\boldsymbol{h}_{i:i+k-1}$ is the higher semantic representation from the last layer. By setting different filter sizes, we can obtain a feature map with multi-granularity information. This **multi-granularity information** indicates that the convolutional operation with different filter sizes can capture the different ranges of n-gram information. When a filter traverses each window in the argument from $\boldsymbol{h}_{1:k-1}$ to $\boldsymbol{h}_{n-k+1:n}$, we obtain the output of the feature map corresponding to the filter $\boldsymbol{f}_m$:

$$\boldsymbol{c}^m = [\boldsymbol{c}^m_1, \boldsymbol{c}^m_2, \ldots, \boldsymbol{c}^m_{n-k+1}]. \tag{9}$$

Here, $\boldsymbol{c}^m \in \mathbb{R}^{n-k+1}$ has different dimensions for different arguments because the arguments are different from each other in length n. The pooling operation captures the most important feature for each feature map, and handles variable sentence lengths naturally.

2) DYNAMIC CHUNK-BASED MAX POOLING LAYER

In general, we apply a standard max pooling operation over [30], [31] over $\boldsymbol{c}^m$ and choose the maximum value $max\{\boldsymbol{c}^m\}$ as the most important feature of the filter $\boldsymbol{f}_m$. However, most of the information of discourse arguments is lost, including the word order. Kalchbrenner *et al.* [13] proposed taking the top-k maximum values over $\boldsymbol{c}^m$ to retain more important information, but the word order was still missing because the features selected by k-max pooling are likely to be concentrated in one part of the argument. Hu *et al.* [23] designed a max pooling over every two-unit, but these might have some redundant information. Chen *et al.* [25] presented a dynamic multi-pooling strategy to split each feature map into three parts by the event triggers and arguments,[2] and choose the maximum of each split part. Zhang *et al.* [24] argued that from the perspective of shallow structures, one sentence is

[2]Notice that "argument" in event extraction is different from "discourse argument".

organized by subject-verb-object (English and Chinese word order), and a sentence can be described as a sequence of a noun phrase, a verb phrase, and an adjective phrase from the perspective of deep structures. Therefore, they presented a chunk-based convolutional neural network to learn sentence semantic representations that retain the sentence structure to some extent, but their chunk size was predefined.

Inspired by these studies, we devise a dynamic chunk-based max pooling operation to divide the argument into several segments (also called chunks) according to the length of the argument and structure of our network and then select the maximum value of each chunk to retain more effective information such as word order of the arguments. Therefore, our **dynamic chunk** represents that the segments of the feature map obtained by the convolution layer, which is divided evenly into ***seg*** segments. The number of chunks ***seg*** is dynamically calculated by Eq.(10). Notice that if the feature $\boldsymbol{c}_i^m$ is not divisible by ***seg***, the last chuck has the size of the modulus.

$$\boldsymbol{seg} = max(\boldsymbol{seg}_{top}, \lfloor \frac{L-l}{L} \rfloor \cdot \boldsymbol{s}). \tag{10}$$

where $\boldsymbol{seg}_{top}$ is the fixed pooling parameter for the topmost convolution layer, L is the number of total layers in the CNN module of the framework, l is the number of the current layer, and $\boldsymbol{s}$ is the length of a sentence. The feature map $\boldsymbol{c}^m$ of one argument is divided into ***seg*** parts equally, and we can take the maximum value from each part. If there are $\boldsymbol{f}_M$ filters, the output of the pooling layer is a vector in the $\boldsymbol{f}_M \times \boldsymbol{seg}$ dimension.

$$\begin{aligned} \boldsymbol{c}_{seg}^m &= ChunkMax\{[\boldsymbol{c}_1^m, \boldsymbol{c}_2^m, \ldots, \boldsymbol{c}_{n-k+1}^m]\} \\ &= [\boldsymbol{c}_{seg1}^m, \boldsymbol{c}_{seg2}^m, \ldots, \boldsymbol{c}_{segn}^m]. \end{aligned} \tag{11}$$

D. CLASSIFIER OUTPUT

After reshaping the output of the pooling layer, the vectors are fed into a full connection hidden layer with non-linear activation to obtain the more abstractive representations, and then connected to the output layer. For the task of classification, the outputs are probabilities of different classes, which are calculated by a softmax function.

E. MODEL TRAINING

Given a training corpus which contains n instances $\{(\boldsymbol{x}, \boldsymbol{y})\}_{r=1}^n$, $(\boldsymbol{x}, \boldsymbol{y})$ denotes an arguments pair and its label. We employ the cross-entropy error to assess how well the predicted relation represents the real relations, defined as:

$$L(\hat{\boldsymbol{y}}, \boldsymbol{y}) = -\sum_{j=1}^{C} \boldsymbol{y}_j \log(Pr(\hat{\boldsymbol{y}}_j)). \tag{12}$$

where $Pr(\hat{\boldsymbol{y}}_j)$ is the predicted probabilities of the j-th label, C is the class number.

To minimize the objective, we use stochastic gradient descent with the diagonal variant of AdaGrad with mini-batches. The parameter update for the i-th parameter $\boldsymbol{\theta}_{t,i}$ at step $\boldsymbol{t}$ is as follows:

$$\boldsymbol{\theta}_{t,i} = \boldsymbol{\theta}_{t-1,i} - \frac{\boldsymbol{\alpha}}{\sqrt{\sum_{\tau=1}^{t} g_{\tau,i}^2}} \boldsymbol{g}_{t,i} \tag{13}$$

where $\boldsymbol{\alpha}$ is the initial learning rate and $\boldsymbol{g}_\tau \in \mathbb{R}^{\theta_{\tau,i}}$ is the gradient at step τ for parameter $\boldsymbol{\theta}_{\tau,i}$.

TABLE 1. The statistics of implicit discourse relations in the PDTB.

Relation	Train	Dev.	Test
Comparison	1842	188	144
Contingency	3139	272	266
Expansion	6658	635	537
Temporal	579	47	55
Total	12218	1142	1002

III. EXPERIMENTAL PREPARATION

A. DATASETS

To evaluate our proposed model, we adopt two corpora: the PDTB and HIT-CDTB datasets are annotated for discourse relation recognition.

Penn Discourse TreeBank[3] (PDTB) [32] is one of the largest hand-annotated discourse relation corpora. It contains approximately 40,600 relations, which are manually annotated from 2,312 Wall Street Journal (WSJ) articles. Discourse relations are organized into a 3-level hierarchy, i.e., class, type and sub-types. Our experiments are conducted on the four top-level classes of PDTB as in previous studies [9], [17], [18], i.e., Comparison (Comp.), Contingency (Cont.), Expansion (Exp.) and Temporal (Temp.). Following conventional data splitting, we select Sections 2-20 as the training set, Sections 21-22 as the testing set, and Sections 0-1 as the development set. Notice that the data preparation of the Expansion relation follows Rutherford and Xue [33], and it is different from Ji and Eisenstein [34] in which they merge the EntRel relation into the Expansion relation. Table 1 presents the data distribution of top-level discourse relations in the PDTB.

HIT-CDTB[4] is a Chinese discourse relation corpus that was annotated by the Harbin Institute of Technology in China [35]. They analysed the differences between Chinese and English, and proposed a modified approach with the relevant Chinese linguistics theory. According to the physical structural units, the discourses were divided into three categories: sentence group, complex sentence, and sub-clause.

There are six relations, namely, Temporal, Causal, Conditional, Comparison, Expansion, and Coordinating in HIT-CDTB, which annotate 1,096 articles from the OntoNotes4.0. Due to focusing on implicit discourse relation recognition, we demonstrate the distribution of six implicit relations in HIT-CDTB, as shown in Figure 2.

In addition, the numbers of temporal and conditional relations are too low. Therefore, we use the remaining relations as experimental data in Table 2.

[3]More details can be found at `https://www.seas.upenn.edu/~pdtb/`.

[4]The details can be found at `http://ir.hit.edu.cn/hit-cdtb/`.

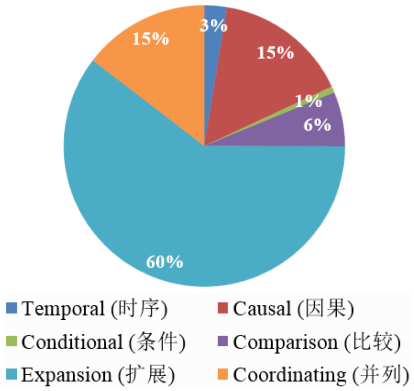


FIGURE 2. The distribution of six Chinese implicit relations.

TABLE 2. The statistics of four implicit discourse relations in HIT-CDTB.

Relation	Train	Dev.	Test
Causal	1068	305	153
Comparison	431	123	63
Expansion	4161	1189	595
Coordinating	1005	287	144
Total	6665	1904	955

B. EXPERIMENTAL SETTINGS

To evaluate our model and compare it with the previous work, we adopt two experimental settings: 1) a multi-class classification, and 2) four separate one-vs-other binary classifications. The first setting is for observing the overall performance on this task, which is more natural in realistic settings. The second setting is for solving the problem of unbalanced data to some extent, where each top-level class is against the other three discourse relation classes. For the second setting, we use an equal number of positive and negative instances as training data in each relation because each of the relations except Expansion is infrequent. Negative instances are chosen randomly from the training set. The testing set and development set are kept in the natural state.

C. PARAMETER SETTINGS

We first pre-process the corpora, such as converting the tokens in PDTB to lowercase and Chinese word segmentation by Jieba[5] for HIT-CDTB. We choose the pre-trained word embeddings for both English and Chinese. We select GloVe [36] published by the NLP group of Standford University for English word embedding; we utilize word2vec to train the Chinese word embedding on Wikipedia and ignore the words that appear less than 5 times in the vocabulary. If there are any words not in the pre-trained vectors, they are randomly generated in $[-1, 1]$, and the dimension of these words is set to 50 as well as the pre-trained word embeddings. And then we statistic the length of all arguments and set it as 80, and we apply truncating or zero-padding operation when necessary. We adopt the filters of 3, 4, 5 with 100 feature maps each for CNN module, which can obtain different range of n-gram features, i.e. "multi-granularity" information.

[5]https://pypi.org/project/jieba/

TABLE 3. Hyper-parameters for our DC-BCNN model.

Hyper-parameters	Value
Initial learning rate	$\alpha = 0.01$
Region sizes	region size = (3, 4, 5)
Stride of convolution operation	stride=1
Minibatch size	mini-batch = 32
Dropout rate	$\rho = 0.1$
The weight of l_2	$\lambda = 0.0001$

The length of the intermediate representation is also set to 50. The other parameters are initialized by random sampling from a uniform distribution in $[-0.1, 0.1]$. Here, we also employ a simple grid search to set the hyper-parameters and give some final settings as shown in Table 3.

D. EVALUATION MEASURES

Furthermore, we follow the criteria of previous works for evaluation with several different metrics, including precision (P), recall (R), and their harmonic mean (F_1). We also report accuracy for a direct comparison with the state-of-the-art models.

E. THE COMPARATIVE SYSTEMS

On the PDTB. We compare against the published results of the following competitive systems, which are from traditional feature-based approaches to various neural network-based models.

1) FEATURE-BASED MODELS

- **Rutherford2014** [9] employed Brown cluster pairs to represent discourse relations and incorporated coreference patterns to identify senses of implicit discourse relations in naturally occurring text.
- **Ji2015** [34] utilized two recursive neural networks on the syntactic parse tree to induce argument representation and entity spans.

2) CONVOLUTIONAL NEURAL NETWORKS

- **Zhang2015** [10] proposed a simplified neural network containing only one hidden layer and three different pooling operations (max, min, average) for the task.
- **Qin2016a** [16] integrated a CNN for sentence modeling and a Collaborative Gated Neural Network (CGNN) for feature transformation into the classification task.
- **Zhang2016** [38] employed semantic memory to encode semantic knowledge of words in arguments, and used an attention model to retrieve the relevant information of argument representations.

3) RECURRENT NEURAL NETWORKS

- **Liu2016** [17] designed neural networks with multi-level attention (NNMA) and selected the important words for recognizing discourse relations. Here, we select the models with two and three levels of attention as baselines.
- **Lan2017** [39] presented two types of representation learning at the same time: 1) an attention-based

neural network, which conducted the representation with interactions; 2) multi-task learning, which leveraged knowledge from the auxiliary task to enhance the performance.

On the HIT-CDTB. Due to little research on Chinese implicit discourse relation recognition, we only choose the following three models as baselines to evaluate the performance of our model on the Chinese corpus. Notice that we reimplemented these models on the Chinese corpus we used.

- **Zhang2013** [35] combined lexical, syntactic and semantic features in a supervised model to classify implicit discourse relations. They trained the maximum entropy (ME) and support vector machine (SVM) models to classify the relations. We only utilize the SVM model as a baseline with their linguistic features.
- **Qin2016b** [40] utilized a convolutional neural network model to determine the senses for both English and Chinese tasks.
- **Rönnqvist2017** [41] proposed the first attention-based recurrent neural sense classifier, specifically developed for Chinese implicit discourse relations.

In addition, we utilize the following ablation models to verify the effectiveness of each component on both English and Chinese datasets.

- **SVM**: A support vector machine classifier for discourse relation recognition by using the human-designed lexical and syntactic features. We adopt the pre-trained word embeddings as inputs of a conventional multi-class SVM [35], [35].
- **CNN**: Adopting pre-trained word embeddings to replace the original words in discourse arguments, we utilize convolutional operation to extract the semantic features of different aspects in arguments by different filter sizes [43].
- **LSTM**: We encode two discourse arguments by LSTMs. Then, we concatenate the two representations and feed them to the full connection hidden layer as the input of the softmax classifier.
- **BiLSTM**: Based on **LSTM**, we consider the bidirectional context information and use BiLSTMs to encode two discourse arguments.
- **BiLSTM + CNN with max pooling (BCM)**: Based on the **BiLSTM** model, we obtain the semantic representations of the arguments with the contextual information, and then we adopt the convolutional neural network to capture different granularity features with a standard max pooling operation.
- **BiLSTM + CNN with k-max pooling (BCKM)**: Different from **BCM** approach, we apply k-max pooling instead of the general max pooling to avoid losing more information.

IV. RESULTS AND DISCUSSION

To verify the effectiveness of our proposed model, we conduct a series of comparison experiments on state-of-the-art systems and the ablation models from different aspects, and then we present an in-depth analysis of different parameters.

TABLE 4. Performance of multiple binary classification on PDTB in terms of F_1 score.

Model	Comp.	Cont.	Exp.	Temp.
Rutherford2014	39.70%	54.42%	70.23%	28.69%
Ji2015	35.93%	52.78%	-	27.63%
Zhang2015	34.22%	52.04%	69.59%	30.54%
Qin2016a	41.55%	57.32%	71.50%	35.43%
Zhang2016	34.82%	53.70%	70.95%	36.16%
Liu2016 (two levels)	36.70%	54.48%	70.43%	38.84%
Liu2016 (three levels)	39.86%	53.69%	69.71%	37.61%
Lan2017	40.73%	58.96%	72.47%	38.50%
Our DC-BCNN	39.84%	56.31%	71.59%	37.76%

TABLE 5. Performance of multi-class classification on PDTB in terms of accuracy (Acc.) and F_1 score.

Model	F_1	Acc.
Ji2015	38.52%	36.98%
Liu2016 (two levels)	46.29%	57.17%
Liu2016 (three levels)	44.95%	57.57%
Lan2017	47.80%	57.39%
Our DC-BCNN	46.39%	57.69%

A. COMPARISON WITH STATE-OF-THE-ART SYSTEMS

On the PDTB. Tables 4 and 5 show the overall performance in detail. For binary classification, the F_1 score is adopted to evaluate the performance on each class. For multi-class classification, the F_1 score and accuracy are used as evaluation metrics. We make the following observations with respect to the binary classification:

- Overall, the performance based on neural network models is significantly better than that of feature-based models, which indicates that the artificially well-designed features are not sufficient for implicit discourse relation classification, and automatically extracting features based on neural networks can capture richer semantic clues. Among them, the BiLSTM-based approaches outperform the CNN-based models because the two BiLSTM-based models leverage other complicated strategies to capture more information, such as multi-level attention in Liu2016 and multi-task learning in Lan2017.
- For each relation, the F_1 scores of the Temporal relation are the lowest in all models, which is reasonable since it accounts for the smallest number of instances (only 5%) in the corpus. With the increase in the number of instances in different relations, the F_1 scores also rise. This proves that the corpus is also crucial to implicit discourse relation recognition.
- Although our DC-BCNN model does not improve the F_1 score compared with Lan2017, our model obtains the comparable scores. Especially, Lan2017 achieves the state-of-the-art performance in recognizing the Contingency relation for the following reasons: (1) some discourse arguments may have confusing information,

TABLE 6. Comparisons with the state-of-the-art models on HIT-CDTB.

Model	P	R	F_1
Zhang2013	32.57%	51.82%	40.43%
Qin2016b	43.87%	64.09%	52.08%
Rönnqvist2017	45.06%	66.35%	53.67%
Our DC-BCNN	48.51%	67.10%	56.31%

which could require more in-depth analysis by different strategies; (2) our DC-BCNN framework integrates the advantages of BiLSTM and CNN models, but some implicit discourse argument pairs need more extensive knowledge for inference. Similar results are obtained in the Comparison relation.

With respect to the multi-class classification, we make the following observations:

- Ji2015 achieves the worst performance on both F_1 score and accuracy. Because it is mainly dependent on integrating a syntactic parse tree, which may create some error propagation problems and ignores the deeper semantic features of discourse arguments.
- Both the Liu2016 and Lan2017 models are based on a recurrent neural network for the task. Although Lan2017 achieves 1.51% more than Liu2016 (two levels) and 1.41% more than our DC-BCNN on the F_1 score, we all achieve the comparable performance on accuracy. However, Lan2017 adopts an attention-based approach to obtain the representation with important information and utilizes multi-task learning to leverage knowledge from the auxiliary task. It integrates considerable information to enhance the performance, which could be complicated and computationally intensive. In addition, the F_1 score of the Liu2016 (two levels) model is higher than that of three levels of attention (1.34%). It indicates that the attention mechanism is useful, but more attention may create the over-fitting problem due to more parameters.
- Despite the fact that the performance of our DC-BCNN model does not improve the F_1 score compared with Lan2017, our model neither utilizes more mechanisms to capture the significant information nor extends the extend knowledge by multi-task learning from the auxiliary task. It illustrates that our model utilizes the advantages of both BiLSTM and CNN to reserve more semantic features; additionally, the convolution operation in our model extracts multi-granularity features from the perspective of n-gram, and the chunk-based max pooling strategy avoids losing the location information by preserving the maximums in pooling windows.

On the HIT-CDTB. We choose the following systems as our baselines to validate the performance of our proposed model in Chinese. Table 6 demonstrates the precision, recall and F_1 score of multi-class classification. The observations we make are as follows:

- As shown in Table 6, we find that all the neural network-based models outperform Zhang2013 obviously. Because Zhang2013 only adopts the following features to train SVM: core verbs, polarity feature, dependent syntax feature, unigram and bigram. These features have limitations that are not enough to represent the semantics of discourse arguments. It is worth mentioning that Zhang2013 has high precision values and low recall values in the relation classifications except for Expansion. It illustrates that the unbalanced instances make the model tend to divide the test instances as the Expansion relation, which leads to the recall increasing.
- Referring to the previous studies [40], [41], we simply reproduce the Qin2016b and Rönnqvist2017 models that adopted CNN-based and RNN-base approaches. Particularly, they gain improvements over Zhang2013 by 11.30%, 12.27%, 11.65% and 12.49%, 14.53%, 13.24% on precision, recall, and F_1, respectively. The results strongly demonstrate that the neural network-based models not only address the data sparsity problem to some extent, but also capture deeper semantic information to infer the implicit discourse relations.
- Our DC-BCNN model obtains slightly better results over Rönnqvist2017 by 3.45%, 2.64% on precision and F_1, respectively. We conjecture that the main reason for this lies in the advantages of both BiLSTM and CNN, which may provide more evidence for discourse relation recognition. Among them, CNN with different sizes of filters tends to retain more multi-granularity information. In addition, a dynamic chunk-based max pooling operation we designed fixes an issue where the standard pooling operation lost the word order information. Therefore, our proposed model could be well adapted to the HIT-CDTB, which has three categories: sentence group, complex sentence, and sub-clause. The word order information is crucial to these three discourse units. Notice that we perform the significance test for these improvements, and they are significant under one-tailed t-test ($p < 0.05$).

TABLE 7. Comparisons with the ablation models on PDTB.

Model	Comp.	Cont.	Exp.	Temp.
SVM	32.28%	48.87%	62.07%	24.61%
CNN	30.50%	51.73%	69.24%	30.19%
LSTM	32.95%	43.38%	68.10%	30.80%
BiLSTM	34.01%	44.68%	68.53%	31.27%
BCM	39.15%	53.44%	68.85%	33.79%
BCKM	39.42%	54.77%	69.03%	35.18%
Our DC-BCNN	39.84%	56.31%	71.59%	37.76%

B. COMPARISON WITH THE ABLATION MODELS

To evaluate the effectiveness of each part in our proposed model, we take six ablation models to compare with our DC-BCNN.

On the PDTB. Seen from Table 7, we make the following observations:

- Overall, the performance of the conventional SVM method is lower than that of other neural network-based

TABLE 8. Comparisons with the ablation models on HIT-CDTB.

Model	Causal	Comparison	Expansion	Coordinating
SVM	27.15%	18.06%	67.22%	43.38%
CNN	33.80%	19.46%	75.65%	52.90%
LSTM	34.93%	23.17%	75.37%	52.81%
BiLSTM	35.87%	25.41%	77.09%	54.13%
BCM	37.29%	26.87%	79.45%	57.39%
BCKM	37.95%	27.06%	81.07%	57.74%
Our DC-BCNN	38.18%	28.74%	84.86%	58.99%

approaches in the first four rows. It indicates that the latter type of model can automatically capture more semantic features and address the data sparsity problem to some extent. We can see that BiLSTM achieves slightly better performance than LSTM, which is consistent with previous work, as BiLSTM considers the forward and backward direction contextual information, while LSTM only considers the forward information. Additionally, the F_1 score of the CNN is 7.05% better than that of the BiLSTM on Contingency, which may be because this relation requires inference from multiple local information, and CNN provides the capacity to model local word sequences (via convolution operations) which is effective for Contingency.

- Compared with LSTM and BiLSTM, BCM and BCKM models achieve much better performance. The performance of BCKM is slightly better than that of BCM overall, probably because k-max pooling in BCKM obtains more effective information. In particular, the F_1 score of BCKM is higher than that of BiLSTM on Comparison, which gains a 5.31% improvement. BCKM also gains 2.35% improvement compared with the best CNN in the last group on Expansion. It proves that (1) the contextual clues and local information are both important for our task; and (2) more selected features (via k-max pooling operation) are obviously useful, especially Contingency relation and Expansion relation, which could need to understand sufficiently.
- Our DC-BCNN model achieves the best performance on the F_1 score. This illustrates that our model acquires the sentence-level (via BiLSTM preserving history and future information) and multi-granularity combined information (via setting the sizes of the filters), which are good for recognizing implicit discourse relations. Different from BCKM which utilizes the subsequence of the original features to obtain the relevant location, we devise a dynamic chunk-based max pooling strategy to retain the location information of the whole argument.

On the HIT-CDTB: Seen from Table 8, we obtain the following observations:

- Similarly, all ablation models have the worst performance in Comparison relation due to the lack of data in HIT-CDTB, which is also consistent with the previous experiment in English. It illustrates that the more instances there are, the better the performance the models will obtain, as with our DC-BCNN model.
- Additionally, similar to the results of ablation experiments on the PDTB, we find that the performance on the F_1 score of neural network-based methods is better than that of the traditional SVM model. The reasons are as follows: on the one hand, the lexical and syntax features used in SVM are mostly separate words or phrases, which may lead to the data sparsity problem; on the other hand, although the feature could be represented by word embedding, the manually annotated features are not enough to represent the discourse arguments.
- Although LSTM and BiLSTM achieve better performance than CNN overall, CNN achieves comparable performance on Expansion and Coordinating relations. It indicates that these two relations require deeper features from the aspect of multi-granularity, which is extracted by convolution operations with different filters.
- BCM and BCKM have better experimental results than simple neural networks. Specifically, they are 3.26% and 3.61% better than BiLSTM on the F_1 score of Coordinating. The performance of BCKM is higher than that of BiLSTM on the other relations, which improves 2.08%, 1.65% and 3.98% in F_1 score, respectively. This demonstrates that the semantic information of different units in discourse arguments is useful for recognizing relations, and it is also beneficial for choosing more features by k-max pooling.
- In addition, our DC-BCNN model outperforms the above approaches. Specifically, DC-BCNN improves 1.68%, 3.79%, and 1.25% compared with the BCKM model on F_1 score of Comparison, Expansion and Coordinating relations, respectively. This not only illustrates that the location (spatial) information, which is captured by our dynamic chunk-based max pooling operation, is an important clue in Chinese for identifying the relations, but also proves the effectiveness of our proposed model. The performance of our model in Causal is not effectively improved, which may be because the Causal relation contains some identifiable patterns (e.g., word pairs with polarity), and the location information has little effect on it. Here, the improvements are significant under one-tailed t-test ($p < 0.05$).

V. RELATED WORK

Along with the increasing requirement, many approaches have been explored for implicit discourse relation recognition. Traditional feature-based methods rely on artificial and shallow features, such as polarity tags, Levin verb classes, verb phrases and word position, which require considerable human effort and are prone to error propagation and data sparsity problems. Recent neural network-based models achieve better performance, and can be roughly divided into the following aspects:

A. CONVOLUTIONAL NEURAL NETWORKS

1) GENERAL POOLING OPERATION

More recently, neural network-based approaches have attracted much attention in the field of NLP.

The convolutional neural networks have demonstrated success in implicit discourse relation recognition. Zhang *et al.* [10] adopted a pure neural network with three different pooling operations for learning shallow representations in the task. In Qin *et al.* [16] (2016a), a CNN with max pooling strategy was used for sentence modelling and a collaborative gated neural network (GCNN) was proposed for feature transformation. These models detect the semantic features by different convolutional operations as local information but are not effective for the long-term dependency of discourse arguments.

Meanwhile, the general pooling operations also have shortages. For example, the conventional max pooling operation loses the location information by preserving the maximum in pooling window; and only with the maximum feature could miss the other useful clues for implicit discourse relation recognition.

2) IMPROVED POOLING OPERATION

CNN rises from the field of image recognition, which also faces the problem of preserving spatial information. There are many relevant studies that bring us new inspiration. Malinowski and Fritz [48] proposed a flexible parameterization of the spatial pooling strategy and learned the pooling regions together with the classifier, which could work with both sum-pooling and max-pooling. Zhai *et al.* [49] presented a novel pooling strategy with stochastic spatial sampling (S3Pool), where the regular downsampling is replaced by a more general stochastic version. And the general stochasticity as a strong regularizer could be seen as doing implicit data augmentation by introducing distortions in the feature maps. Zheng *et al.* [50] extracted displacement information that recorded the location of the maximal values in the max pooling operation, and combined them with the features resulting from max-pooling, which tackled the issues with structural deformations of max pooling in handwritten text recognition. However, the pooling operations in the fields of image and natural text have essential differences. We could not employ these pooling operations directly in natural language tasks.

Inspired by more pooling operations in image field, many researches focused on the pooling issues of natural text. For modelling sentence task, Kalchbrenner *et al.* [13] designed a dynamic k-max pooling to select the top-k maximum values from each feature map, which could retain more valuable information. They claimed that the top-k maximum values is actually a subsequence of the original features, which could obtain the relevant location cues. Chen *et al.* [25] split each feature map into three parts according to event triggers and arguments. They kept the max value of each split part to capture more valuable information. Zhang *et al.* [24] proposed a bilingually-constrained chunk-based convolutional neural network for event extraction task.

Although these improved pooling operations solved the issues of location clues to some extend, the dynamic k-max pooling only kept the sequence information of ''subsequence'' without the entire features; the chunk-based pooling layer could retain considerable information, the chunk size was predefined, which may cause some crucial features to be lost;considered the characteristic of specific task, the dynamic multi-pooling [25] could not be directly applied to other tasks.

B. RECURRENT NEURAL NETWORKS

Another popular choice of network is the recurrent neural network including its variants (e.g., LSTM, BiLSTM and GRU) [17], [18], [44]–[46]. Liu *et al.* [17] imitated the repeated reading strategy, and further proposed the neural networks with multi-level attention. This model encoded discourse argument by BiLSTM, combining the attention mechanism and external memories to gradually fix the attention on some specific words for identifying discourse relations. Chen *et al.* [18] encoded the discourse argument to its positional representation via BiLSTM, and employed a gated relevance network to capture the semantic interaction between the arguments, which overcomes the semantic gap. Lan *et al.* [39] analysed the discourse argument from LSTM to attention neural network, and further proposed multi-task learning framework to address the implicit discourse relation recognition with the aid of large amount of unlabelled data.

The RNN and its variants could solve the long-term dependency to some extent. Although the ones using attention or memory mechanisms captured more specific important information, they are poor at dealing with the local features.

C. HYBRID NEURAL NETWORK

Different recurrent neural networks and convolutional neural networks have won tremendous success in our task, where either RNN or CNN has its advantages and disadvantages. For instances, RNN is good at capturing sequence feature, particularly the long-distance dependency, but could not obtain the local features; the capability of CNN is extracting local information of different linguistic units, but it is poor at the long-term dependencies of sentences. Considering the merits of both, many researchers tend to integrate RNN and CNN into one architecture to obtain the global and local features. Qin *et al.*(2016b) [40] presented a combination of a BiLSTM and CNNs. They constructed character-based word representations by transforming character embeddings with CNN and BiLSTM layers. Another CNN layer was used to extract an argument representation from a sequence of words. Zhang *et al.* [51] combined BiLSTM-CNN to extract series of higher-level phrase representations for relation classification. In addition, Guo *et al.* [52] extracted the discriminative local interactions between amino-acid residues by 2D CNNs, and further capture long-distance interactions between amino-acid residues by bidirectional gated recurrent units or BiLSTM, which improved the protein secondary structure prediction. They [53] also proposed a novel DeepACLSTM model, which applied asymmetric convolutional neural networks combined with BiLSTM to predict protein secondary structure.

In summary, there are two ways of combining RNN and CNN. One is CNN-RNN framework which is suitable for speech recognition [54], [55], biomedical engineering [52], [53] and other fields. Generally, these studies first extract the local features and then construct the long-term dependencies, due to the characteristics of their inputs (i.e., images). Another is RNN-CNN architecture for natural text tasks [40], [51]. It utilizes RNN to encode the contextual information of text and then captures local clues from consecutive context, which is also in line with the process of people's cognitive understanding.

As mentioned above, various methods have been proposed for English (PDTB), while the Chinese task has received little attention in the literature. Liu *et al.* [47] utilized an attention-based neural network to represent arguments and employed an external memory network to preserve crucial information for Chinese task. We argue that Chinese needs word segmentation and other pre-processing, which is more sensitive to multi-granularity information.

Inspired by the relevant work, we integrate BiLSTM models that represent discourse arguments with contextual information and CNNs that capture the semantic features from the wider ranges of n-gram. We devise a dynamic chunk-based max pooling operation, which dynamically divides the argument into several segments to capture more maximum values for retaining as much information as possible. Our entire framework can address the mentioned disadvantages to some extent.

VI. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel **D**ynamic **C**hunk based Max Pooling **B**iLSTM-**CNN** model (DC-BCNN) for implicit discourse relation recognition. We integrate the merits of BiLSTM and CNN and further design dynamic chunk-based max pooling to capture crucial information in different granularity feature maps. Essentially, we are inspired by the reading experience, and people can capture different semantic information from different granularity representations during reading. Thus, they can understand the argument meanings and determine their relation by combining that information. Therefore, after obtaining the semantic representations of two discourse arguments by BiLSTMs, we utilize convolutional operation with different sizes of filters to extract different aspects of semantic features, which captures more information from a wider range of n-grams. Meanwhile, we also devise a dynamic chunk-based max pooling operation to maintain the word order, which can obtain crucial information about each part of the arguments. The experimental results on PDTB and HIT-CDTB corpora both show that our model is effective.

However, we only focus on capturing the inter-sentence information and ignore the wider contexts beyond two arguments. In the future, we plan to not only establish interdependencies between higher hierarchical discourse units, but also exploit external knowledge to effectively improve implicit discourse relation recognition.

REFERENCES

[1] T. Meyer and A. Popescu-Belis, "Using sense-labeled discourse connectives for statistical machine translation," in *Proc. 13th EACL*, Apr. 2012, pp. 129–138.

[2] S. Gerani, Y. Mehdad, G. Carenini, R. T. Ng, and B. Nejat, "Abstractive summarization of product reviews using discourse structure," in *Proc. EMNLP*, Oct. 2014, pp. 1602–1613.

[3] R. Higashinaka, K. Imamura, T. Meguro, C. Miyazaki, N. Kobayashi, H. Sugiyama, T. Hirano, T. Makino, and Y. Matsuo, "Towards an open-domain conversational system fully based on natural language processing," in *Proc. 25th COLING*, Aug. 2014, pp. 928–939.

[4] A. Otsuka, T. Hirano, C. Miyazake, R. Higashinaka, T. Makino, and Y. Matsuo, "Utterance selection using discourse relation filter for chat-oriented dialogue systems," in *Dialogues With Social Robots*. Singapore: Springer, 2017, pp. 355–365.

[5] Y. Zhou and N. Xue, "PDTB-style discourse annotation of Chinese text," in *Proc. 50th ACL*, Jul. 2012, pp. 69–77.

[6] E. Pitler, A. Louis, and A. A. Nenkova, "Automatic sense prediction for implicit discourse relations in text," in *Proc. AFNLP*, Aug. 2009, pp. 683–691.

[7] Z.-M. Zhou, Y. Xu, Z.-Y. Niu, M. Lan, J. Su, and C. L. Tan, "Predicting discourse connectives for implicit discourse relation recognition," in *Proc. 23rd COLING*, Aug. 2010, pp. 1507–1514.

[8] J. Park and C. Cardie, "Improving implicit discourse relation recognition through feature set optimization," in *Proc. 13th Annu. Meeting Special Interest Group Discourse Dialogue*, Jul. 2012, pp. 108–112.

[9] A. Rutherford and N. Xue, "Discovering implicit discourse relations through brown cluster pair representation and coreference patterns," in *Proc. 14th EACL*, Apr. 2014, pp. 645–654.

[10] B. Zhang, J. Su, D. Xiong, Y. Lu, H. Duan, and J. Yao, "Shallow convolutional neural network for implicit discourse relation recognition," in *Proc. EMNLP*, Sep. 2015, pp. 2230–2235.

[11] Y. Liu, S. Li, X. Zhang, and Z. Sui, "Implicit discourse relation classification via multi-task neural networks," in *Proc. 30th AAAI*, Mar. 2016, pp. 2750–2756.

[12] R. Socher, E. H. Huang, J. Pennin, C. D. Manning, and A. Y. Ng, "Dynamic pooling and unfolding recursive autoencoders for paraphrase detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 801–809.

[13] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," Apr. 2014, *arXiv:1404.2188*. [Online]. Available: https://arxiv.org/abs/1404.2188

[14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," Apr. 2014, *arXiv:1404.2188*. [Online]. Available: https://arxiv.org/abs/1404.2188

[15] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. EMNLP*, Oct. 2013, pp. 1631–1642.

[16] L. Qin, Z. Zhang, and H. Zhao, "A stacking gated neural architecture for implicit discourse relation classification," in *Proc. EMNLP*, Nov. 2016, pp. 2263–2270.

[17] Y. Liu and S. Li, "Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention," in *Proc. EMNLP*, Nov. 2016, pp. 1224–1233.

[18] J. Chen, Q. Zhang, P. Liu, X. Qiu, and X. Huang, "Implicit discourse relation detection via a deep architecture with gated relevance network," in *Proc. 54th ACL*, Aug. 2016, pp. 1726–1735.

[19] W. Lei, X. Wang, M. Liu, I. Ilievski, X. He, and M.-Y. Kan, "SWIM: A simple word interaction model for implicit discourse relation recognition," in *Proc. 26th IJCAI*, Aug. 2017, pp. 4026–4032.

[20] A. Rutherford, V. Demberg, and N. Xue, "A systematic study of neural discourse models for implicit discourse relation," in *Proc. 15th Conf. Eur. Chapter Assoc. for Comput. Linguistics*, vol. 1, Apr. 2017, pp. 281–291.

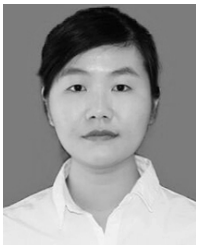
[21] X. Chu, F. Jiang, and N. Xue, "Joint modeling of structure identification and nuclearity recognition in macro Chinese discourse treebank," in *Proc. 27th COLING*, Aug. 2018, pp. 536–546.

[22] F. Jiang, P. Li, X. Chu, Q. Zhu, and G. Zhou, "Recognizing macro Chinese discourse structure on label degeneracy combination model," in *Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput.*, Aug. 2018, pp. 92–104.

[23] B. Hu, Z. Lu, H. Li, and Q. Chen, "Convolutional neural network architectures for matching natural language sentences," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2042–2050.

[24] J. Zhang, D. Zhang, and J. Hao, "Local translation prediction with global sentence representation," in *Proc. 24th IJCAI*, Jun. 2015, pp. 1398–1404.

[25] Y. Chen, L. Xu, K. Liu, D. Zeng, and J. Zhao, "Event extraction via dynamic multi-pooling convolutional neural networks," in *Proc. 53rd Annu. Meeting Assoc. for Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, Jul. 2015, pp. 167–176.
[26] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Feb. 2003.
[27] T. Mikolov, K. Chen, G. Gorrado, and J. Dean, "Efficient estimation of word representations in vector space," Jan. 2013, *arXiv:1301.3781*. [Online]. Available: https://arxiv.org/abs/1301.3781
[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[29] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, "Relation classification via convolutional deep neural network," in *Proc. COLING*, 2014, pp. 2335–2344.
[30] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12 pp. 2493–2537, Aug. 2011.
[31] Y. Kim, "Convolutional neural networks for sentence classification," Aug. 2014, *arXiv:1408.5882*. [Online]. Available: https://arxiv.org/abs/1408.5882
[32] R. Prasad, N. Diesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber, "The penn discourse treebank 2.0," in *Proc. 6th LREC*, May 2008, pp. 1–8.
[33] A. T. Rutherford and N. Xue, "Improving the inference of implicit discourse relations via classifying explicit discourse connectives," in *Proc. NAACL*, May/Jun. 2015, pp. 799–808.
[34] Y. Ji and J. Eisenstein, "One vector is not enough: Entity-augmented distributional semantics for discourse relations," *Trans. Assoc. Comput. Linguistics*, vol. 3, pp. 329–344, Dec. 2014.
[35] M. Zhang, Y. Song, B. Qin, and T. Liu, "Chinese discourse relation recognition," *J. Chin. Inf. Process.*, vol. 27, pp. 51–58, 2013.
[36] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, Oct. 2014, pp. 1532–1543.
[37] E. Yee, E. Chrysikou, and S. Thompson-Schill, "The cognitive neuroscience of semantic memory," in *The Oxford Handbook of Cognitive Neuroscience*. Oxford, U.K.: Oxford Univ. Press, 2014.
[38] B. Zhang, D. Xiong, and J. Su, "Neural discourse relation recognition with semantic memory," Mar. 2016, *arXiv:1603.03873*. [Online]. Available: https://arxiv.org/abs/1603.03873
[39] M. Lan, J. Wang, Y. Wu, Z.-Y. Niu, and H. Wang, "Multi-task attention-based neural networks for implicit discourse relationship representation and identification," in *Proc. EMNLP*, Sep. 2017, pp. 1299–1308.
[40] L. Qin, Z. Zhang, and H. Zhao, "Shallow discourse parsing using convolutional neural network," in *Proc. CoNLL Shared Task*, Aug. 2016, pp. 70–77.
[41] S. Rönnqvist, N. Schenk, and C. Chiarcos, "A recurrent neural model with attention for the recognition of Chinese implicit discourse relations," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, Apr. 2017, pp. 256–262.
[42] H.-H. Huang and H.-H. Chen, "Chinese discourse relation recognition," in *Proc. 5th Int. Joint Conf. Natural Lang. Process.*, 2011, pp. 1442–1446.
[43] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," Oct. 2015, *arXiv:1510.03820*. [Online]. Available: https://arxiv.org/abs/1510.03820
[44] D. Cai and H. Zhao, "Pair-aware neural sentence modeling for implicit discourse relation classification," in *Proc. Int. Conf. Ind., Eng. Other Appl. Appl. Intell. Syst.* Cham, Switzerland: Springer, 2017, pp. 458–466.
[45] F. Guo, R. He, D. Jin, J. Dang, L. Wang, and X. Li, "Implicit discourse relation recognition using neural tensor network with interactive attention and sparse learning," in *Proc. 27th COLING*, Aug. 2018, pp. 547–558.
[46] X. Yue, L. Fu, and X. Wang, "Externally controllable RNN for implicit discourse relation classification," in *Proc. Natural Lang. Process. Chin. Comput.*, 2017, pp. 158–169.
[47] Y. Liu, J. Zhang, and C. Zong, "Memory augmented attention model for chinese implicit discourse relation recognition," in *Proc. 16th China Nat. Conf.*, Oct. 2017, pp. 411–423.
[48] M. Malinowski and M. Fritz, "Learning smooth pooling regions for visual recognition," in *Proc. 24th Brit. Mach. Vis. Conf.*, 2013, pp. 1–11.
[49] S. Zhai, H. Wu, A. Kumar, Y. Chen, Y. Lu, Z. Zhang, and R. Feris, "S3pool: Pooling with stochastic spatial sampling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4970–4978.
[50] Y. Zheng, B. K. Iwana, and S. Uchida, "Mining the displacement of max-pooling for text recognition," *Pattern Recognit.*, vol. 93, pp. 558–569, Sep. 2019.
[51] L. Zhang and F. Xiang, "Relation classification via BiLSTM-CNN," in *Proc. Int. Conf. Data Mining Big Data*. Cham, Switzerland: Springer, 2018, pp. 373–382.
[52] Y. Guo, B. Wang, W. Li, and B. Yang, "Protein secondary structure prediction improved by recurrent neural networks integrated with two-dimensional convolutional neural networks," *J. Bioinf. Comput. Biol.*, vol. 16, no. 5, 2018, Art. no. 1850021.
[53] Y. Guo, W. Li, B. Wang, H. Liu, and D. Zhou, "DeepACLSTM: Deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction," *BMC Bioinf.*, vol. 20, no. 1, p. 341, Dec. 2019.
[54] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Schuller, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5200–5204.
[55] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and recurrent neural networks," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Dec. 2016, pp. 1–4.



FENGYU GUO received the B.E. and M.S. degrees from Xinjiang University, Xinjiang, China, in 2011 and 2014, respectively. She is currently pursuing the Ph.D. degree with the Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, Tianjin, China.

Her current research interests include natural language processing and discourse analysis.

RUIFANG HE received the Ph.D. degree in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 2010.

She is currently an Associate Professor with the College of Intelligence and Computing, Tianjin University, China. Her current research interests include natural language processing, social media mining, and machine learning.

JIANWU DANG received the B.E. and M.E. degrees from Tsinghua University, China, in 1982 and 1984, respectively, and the Ph.D. degree from Shizuoka University, Japan, in 1992.

From 1984 to 1988, he was a Lecturer with Tianjin University. From 1992 to 2001, he was with the ATR Human Information Processing Research Laboratories, Japan. Since 2001, he has been on the Faculty of the School of Information Science, JAIST, as a Professor. He joined the Institute of Communication Parlee (ICP), Center of National Research Scientific, France, as a Research Scientist (first class), from 2002 to 2003. Since 2009, he has been with Tianjin University, Tianjin, China. His research interests include speech production, speech synthesis, and speech cognition. He built a 3D physiological model for speech and swallowing and endeavors to apply the model on clinics.