Tree Framework With BERT Word Embedding for the Recognition of Chinese Implicit Discourse Relations

DAN JIANG AND JIN HE
School of Software Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing 100876, China
Key Laboratory of Trustworthy Distributed Computing and Service, Ministry of Education, Beijing University of Posts and Telecommunications (BUPT), Beijing 100876, China
Corresponding author: Dan Jiang (jiangdan@bupt.edu.cn)

ABSTRACT Currently, discourse relation recognition (DRR), which is not directly marked with connectives, is a challenging task. Traditional approaches for implicit DRR in Chinese have focused on exploring the concepts and features of words; however, these approaches have only yielded slow progress. Moreover, the lack of Chinese labeled data makes it more difficult to complete this task with high accuracy. To address this issue, we propose a novel hybrid DRR model combining a pretrained language model, namely bidirectional encoder representations from transformers (BERT), with recurrent neural networks. We use BERT as a text representation and pretraining model. In addition, we apply a tree structure to the implicit DRR in Chinese to produce hierarchical classes. The 19-class F1 score of our proposed method can reach 74.47% on the HIT-CIR Chinese discourse relation corpus. The attained results showed that the use of BERT and the proposed tree structure forms a novel and precise method that can automatically recognize the implicit relations of Chinese discourse.

INDEX TERMS Discourse relation recognition, deep neural network, tree structure, BERT.

I. INTRODUCTION
Discourse relation recognition (DRR) aims to identify the semantic relation of two sentences or clauses. It is crucial to many text-understanding applications, including question answering [1], sentiment analysis [2], and machine translation [3]. According to the existence of connectives, discourse relations can be divided into two categories: explicit and implicit. Explicit means the discourse connectives explicitly exist between two sentences or clauses; conversely, implicit means there is not an exact connective in the text. Explicit discourse relations can be easily recognized owing to these connectives [4], [5], such as “but” and “so”. However, owing to the lack of strong clues, implicit DRR remains a significant challenge.

Some early approaches to implicit DRR relied heavily on feature engineering, which requires domain knowledge [6], [7]. With the development of deep learning, recent studies on implicit DRR have been successful at applying deep neural network models [8]–[12].

Previous studies on DRR have mainly focused on English [4], [11], Hindi [13], Turkish [14], and Arabic [15]. The depth of the research depends partly on the quantity and quality of the corpora. Two of the most widely used discourse banks are the Rhetorical Structure Theory Discourse Treebank [16] and the Penn Discourse Tree Bank (PDTB) [17]. Based on the theory of PDTB, some studies have constructed datasets of Chinese discourse relations [18], [19]. However, numerous researchers have ignored the differences between Chinese and English. For example, there is no “purpose”, “progressive” or “parallel” in the English semantic relation system. Furthermore, “tense” in English has no corresponding relationship in Chinese. The lack of datasets which can reflect the characteristics of Chinese discourse relations makes research difficult in Chinese DRR tasks. To compensate for this shortcoming, Zhang et al. analyzed the characteristics of Chinese discourse, presented the first Chinese discourse relation taxonomy, and constructed the HIT-CIR Chinese Discourse Relation Corpus (HIT-CDTB) [20].

In this study, we propose a novel hybrid model of implicit DRR in Chinese to combine bidirectional encoder representations from transformers (BERT) with the classification representation of a tree structure.
II. RELATED WORK

A. IMPLICIT DISCOURSE RELATION RECOGNITION IN CHINESE

Implicit DRR has been widely studied in recent years, and can be applied to multiple fields. Owing to the shortage of Chinese discourse datasets, research on Chinese implicit DRR has lagged behind studies on English and other languages.

Referring to the tree structure, connector representation, and relationship classification of the PDTB, Zhou and Xue [18] proposed the Chinese Discourse Treebank (CDTB). The appearance of the CDTB has propelled research into Chinese implicit DRR. Many studies on Chinese DRR have followed English research methods and have used bilingual corpora to solve the problem of insufficient data. Rutherford and Xue [21] proposed a robust neural classifier for non-explicit discourse relations for both English and Chinese, and their model only requires word vectors and a simple feed-forward training procedure. In addition, Xu and Wu [22] utilized an English-Chinese alignment corpus to discover implicit Chinese discourse relations.

Previous studies on Chinese implicit DRR have mainly focused on traditional machine learning methods. Kong and Zhou [7] proposed an end-to-end text analyzer that uses linguistic features such as the context, the vocabulary, and a dependency tree to identify text relations using a maximum entropy classifier. Several recent studies have also adopted neural network methods. Ronnqvist et al. [8] introduced an attention-based Bi-LSTM and demonstrated that by modeling the argument pairs, a joint sequence can outperform word order-agnostic approaches. In addition, Li et al. [9] proposed a neural network-based framework that consists of two hierarchies: a model hierarchy and a feature hierarchy. Finally, Xu et al. [10] proposed a topic tensor network using both sentence and topic-level representations. These studies showed that a neural network achieves good performance on DRR tasks, and the effects of a neural network depend on the training data. However, the CDTB ignores the differences between Chinese and English. Therefore, the proposed use of the HIT-CDTB is significant.
B. BERT
BERT is a language representation model that can be fine-tuned to create state-of-the-art models for a wide range of tasks [23]. The major innovation of BERT is the use of a masked language model and next sentence prediction to capture the representation of words and sentences, respectively. Thus, BERT is good at dealing with tasks related to deep semantic understanding and a sentence relationship. Many researchers have applied a pretrained BERT to improve the performance of their semantic analysis tasks. Lee et al. [24] applied BERT as a word embedding method to integrate a BiLSTM network with an attention mechanism for medical text inferences and extracted the deep semantic information of medical texts. Ohsugi et al. [25] proposed a conversational machine comprehension model using BERT to encode a paragraph independently conditioned with each question and answer in a multiturn context.

BERT is good at dealing with tasks in which the text itself contains answers, when the input sequences is less than 512 tokens and the task does not depend on external knowledge. Shaptala and Didenko [26] described an approach to grammatical error correction using BERT to create encoded representations. Their method increases the system productivity and lowers the processing time while maintaining high accuracy results. Gao et al. [27] implemented three target-dependent variations of BERT for sentiment classification and verified that coupling BERT with complex neural networks for embedded representations does not achieve a significant effect.

Moreover, BERT not only is good at English text analysis, but also is good at analyzing other languages, such as Arabic [28], French [29], and Chinese [30]. Yao et al. [30] utilized an unlabeled clinical corpus to fine-tune the BERT language model before training the text classifier and achieved better results than other deep learning models and other state-of-the-art methods.

Regarding the DRR input, pairs of sentences or paragraphs are applied, and the deep semantic and relationship information comes from the text itself. Thus, the Chinese implicit DRR tasks meet the required characteristics achieved by BERT.

III. MODEL
For Chinese implicit DRR, in this study the relation labels are first organized into relation trees according to the level of the relations, and a model based on BERT, BiLSTM, and a tree structure is then applied to complete this task.

Figure 1 shows the model structure proposed in this study. To obtain the relationship between the front and the back sentences, the dataset is divided into two groups, A and B, where A represents the first sentence, and B represents the second sentence. A and B relate the content through a text relationship and are used for fine-tuning through the BERT model to obtain sentence vectors. Through the fusion layer, the vectors are then merged into the BiLSTM model to obtain the relationship features of the two sentences. After feature screening through BiLSTM, the final features are sent into the relationship tree, and the final relation is determined through the tree structure.

A. BERT WORD EMBEDDING
The first task of traditional text classification model based on Recurrent Neural Network (RNN) is to obtain the word vector according to the pre-training of the text based on the dictionary applied. However, a traditional language model obtains the next word with the largest probability after the probability multiplication of the given sequence from beginning to end. Thus, it is difficult to extract the feature of the word-level relationship. Google first proposed the BERT model in 2018. The excellent performance of this model in the NLP field can provide a better solution to text relationship feature extraction. In this study, the BERT model was introduced into the overall structure to obtain the vector of a sentence or clause.

The transformer structure of the BERT model itself contains a self-attention layer, as shown in Figure 2. The self-attention layer can extract the relation weight of the word vector in the text, and thus it can fully obtain the context feature of the text. After passing through this layer, each word notices other words in this discourse, and it relates the semantic features of other words according to their weight.

Considering the differences between English and Chinese in text pretraining process, we adapt whole word masking instead of Chinese characters masking, that the masked tokens in a whole word will be masked altogether. The text semantic can be obtained more accurately in training, when the whole word is token as a unit.
B. FEATURE FUSION
In the DRR processing of group data, we applied BERT to obtain two sentence vector matrices \( W_a \) and \( W_b \). \( W_a = (w_1, w_2, \ldots, w_l), \) \( W_b = (w'_1, w'_2, \ldots, w'_l), \) \( W_{a,b} \in \mathbb{R}^{l \times d_w} \), where \( l \) is the max length of sentences, and \( d_w \) is the size of word vectors. The two sentence vectors are fused using (1).

\[
W_n = F(W_a, W_b) = (w_1, w_2, \ldots, w_l, w'_1, w'_2, \ldots, w'_l). \tag{1}
\]

The two vectors are concatenated without losing feature information and formed a new vector set to represent the two groups of sentences.

To obtain and filter the features of the relationship between the two parts, the BiLSTM model is applied, and it can filter the features of the discourse and obtain a low-dimensional vector to reduce the complexity of the model. Other RNN models, such as BGRU [31], can also extract text features in the discourse semantic analysis task. However, in order to compare our results with the baseline models, which applied LSTM or BiLSTM as their main network, we choose BiLSTM as the feature extraction network of our model.

Through feature fusion and the BiLSTM layer, the vectors can be well connected with the relationship tree structure to complete the relationship classification task. The entire output of BiLSTM layer \( H_n = (h_1, h_2, \ldots, h_n) \) can be obtained as in (2).

\[
H_n = f_{\text{LSTM}}(W_n). \tag{2}
\]

C. TREE STRUCTURE
Each group of data contains two parts (sentences or clauses). The label represents the semantic relationship between the two parts, which is hierarchical. The relationship structure conforms to the characteristics of a tree structure. Therefore, a relationship tree can be constructed. Each primary class is a root node, and the final class is a leaf node. In the example of Figure 3, the label of these two sentences is 2-1-1, which means the relation of these two sentences is “Direct causation” in “Causation” and can be subdivided into “Reason before” as the tertiary class.

The tree structure of this example is shown in Figure 3, and the root node represents the primary relation “Causation”. Causation can be divided into several level-2 classes, and each level-2 class can be divided into two level-3 classes (in this example, the level-2 class “Direct causation” is divided into “Reason before” and “Result before”).

We trained each inner node \( i \) with a learned filter \( w_i \) and a bias \( b_i \) using (3).

\[
P_i(x_i) = \sigma(w_i x_i + b_i), \tag{3}
\]

where \( x_i \) is the input of node \( i \), and \( \sigma \) is the sigmoid logistic function. The tree structure learns a hierarchy of filters that are applied to assign each example to a particular class.

For the classification task based on the tree structure, the weight of each layer is allocated. When the root node is classified incorrectly, the classification result can directly determine the error. When a level-2 or level-3 node is classified incorrectly, corresponding information feedback is created based on the correct weight of the previous layer to improve the classification accuracy.

D. IMPROVED LOSS FUNCTION
The tree structure can be constructed according to the class level. The deeper the tree is, the finer the classification is. A tree structure is applied to obtain the training features of the discourses and complete the final discourse relationship classification through feature selection. The root node of each tree represents the primary class of the discourse relationship, which has the characteristics of a small number of classes and relatively balanced data distribution. Therefore, the cross entropy loss is used to complete the feature selection of this layer, such as through (4).

\[
L_{\text{root}} = -y \log p, \tag{4}
\]

where \( y \) is the indicator variable. If the predicted value is the same as the real value on this node, \( y = 1 \); otherwise, it is zero. In addition, \( p \) represents the prediction probability that the sample belongs to the class on this root node.

After the root node filtering on the first layer, we obtain the primary class of the current sample prediction. After entering the first layer of child nodes, the data distribution of this layer may exhibit an unbalanced phenomenon, that is, the data of a certain type of relationship are relatively concentrated (positive and easy), which will cause the training to deviate. To solve this problem, we introduce the focal loss [32] to downweight easy examples and thus focus the training on the hard examples through (5).

\[
L_i = -\alpha_i(1 - p_i)^{\gamma} \log(p_i), \quad i = 1, \tag{5}
\]

where \( i \) represents the number of the current layer. Lin et al. [31] added a modulating factor \((1 - p_i)^{\gamma}\) to the cross entropy loss, with a tunable focusing parameter \( \gamma \geq 0 \), and \( \alpha_i \in [0, 1] \) balances the importance of positive and negative examples. Because the data distribution features of each child node are similar, the feature screening is the same as...
relationships (6) can be obtained using (5) combined with the parent-child relationship needs to be added to the filter. Because $L_i$ can express the probability of the prediction classes of the nodes in this layer, it can also represent the weight of each node in the layer. Therefore, (6) can be obtained using (5) combined with the parent-child relationships:

$$L_i = L_{i-1} \cdot [-\alpha_i(1 - p_i)\gamma \log(p_i)], \quad i > 1.$$  

The weight value of the previous layer is used to readjust the prediction value of the layer and improve the classification accuracy. Combining the filtering scheme of the root node and leaf nodes, the loss of each node in the tree can be obtained through (7):

$$L_c = \begin{cases} L_{\text{root}}, & i = 0 \\ L_{\text{root}} \cdot L_i, & i > 0 \end{cases}$$  

For the task on the HIT-CDTB, there are four primary classes, with four class trees in total, and the root node of each tree is the primary category. In addition, there are four level classes, that is, $i_{\text{max}} = 3$.

### IV. EXPERIMENTS

#### A. DATASET

The discourses of the HIT-CDTB come from the Chinese news of OntoNotes. There are 16,667 pairs of sentences or clauses in the HIT-CDTB. We set 80% of the dataset as the training data and 20% as the test data. There are four levels of classification of discourse relations, and every basic class (represented as a classification tree) is divided into multiple classes, as shown in Figure 4.

Taking the class in the lower right corner of Figure 4 as an example, there are 6 classes in the second-level and 12 classes in the third-level. Every second-level class has two child nodes, and several third-level classes have child nodes. There are a total of 38 classes in this dataset, and the problem of data imbalance problem is serious. Classifications at all levels interact with each other through the tree structure. Thus, coarse-grained classification can affect the results of fine-grained classification.

#### B. TRAINING

We use Keras from https://keras.io as our deep learning framework and apply the Keras package Keras-Transformers from https://github.com/huggingface/pytorch-transformers as our BERT encoder, which was formerly known as keras bert. All the models were trained on four NVIDIA Titan X GPUs. In general, we set $\alpha_i$ to 0.25, and $r$ to 2 in (5).

We set the output vector of the LSTM layer to 100 dimensions. We trained our model using Adam with gradient clipping. Adam designs an independent adaptive learning rate for different parameters by calculating the estimation of the first-order and second-order moments of the gradient. The dropout layer selects data randomly to prevent overfitting and this makes the model to be much more robust. Furthermore, the dropout rate can affect the time efficiency. Therefore, a reasonable dropout rate is necessary for modeling. We regulated our network with a dropout rate of 0.5 before the output layer. The batch size was set to 32 for the best learning results.

#### C. BASELINES

We benchmarked the following baseline methods for Chinese implicit DRR, which achieved good results.

Rutherford et al. proposed the use of neural network models based on a feedforward and LSTM architecture and systematically studied the effects of varying structures [12]. We chose the proposed LSTM architecture as the baseline for our method because the sequential LSTM model outperforms the feedforward model on a relatively small corpus.

In addition, Ronnqvist et al. released the first attention-based recurrent neural network classifier for Chinese implicit DRR [8].

Liu and Li proposed neural networks with multilevel attention (NNMA) and combined an attention mechanism with external memory to gradually fix the attention on specific words that are helpful in judging the discourse relations [11].

#### V. RESULTS AND DISCUSSION

#### A. OVERALL PERFORMANCE

According to the different classification granularity, the discourse data of the HIT-CDTB can be divided into 4 classes, 19 classes, 32 classes, and 38 classes, respectively, and we take the classification results of 19 classes as an example. As shown in Table 1, the proposed model performs the best overall among all the neural architectures we explored. It outperforms the LSTM network with one hidden layer [12] and performs comparably with the recurrent neural baseline, which uses an attention mechanism for feature concerns [8]. Furthermore, the multilevel attention mechanism
TABLE 1. Results of the 19-class compared with baselines on the HIT-CDTB in terms of the computational elapsed time of one epoch (Time), the accuracy (Acc.), precision (Pre.), recall (Rec.) and average F1 score.

| Model                     | Time/s | Acc.%  | Pre.%  | Rec.%  | F1 score |
|---------------------------|--------|--------|--------|--------|----------|
| LSTM [12]                 | 6.5    | 64.01  | 67.02  | 61.73  | 64.21    |
| BiLSTM + Attention [8]    | 9.8    | 69.59  | 76.14  | 60.83  | 67.34    |
| BiLSTM + multi-level attention [11] | 16.8   | 70.20  | 79.12  | 60.23  | 68.51    |
| Ours                      | 15.7   | 76.07  | 79.60  | 70.58  | 74.47    |

was demonstrated to be appropriate for the semantic extraction of long discourses. However, the NNMA with multilevel attention [11] achieves an inferior F1 value compared with our method because of the excellent performance of our model with the tree structure. From the computational elapsed time of all the experimental models, it can be seen that our model has a relatively complex structure, but the computational elapsed time is not more than that of other models.

We also tested our trained method on the CDTB [18] and compare its results with the baselines. Our model only tested EntRel, Conjunction, Expansion, and Causation, which are the top four most frequent relations in the test set. There is not sufficient data on the other relations to use them for classification.

TABLE 2. Results of 4-class compared with the baselines on the CDTB in terms of the computational elapsed time of one epoch (Time), the accuracy (Acc.) and the microaveraged and macroaveraged F1 score.

| Model                     | Time/s | Acc.%  | micro-F1 | macro-F1 |
|---------------------------|--------|--------|----------|----------|
| LSTM [12]                 | -      | 70.47  | -        | -        |
| BiLSTM + Attention [8]    | -      | 73.01  | 65.30    | 53.50    |
| BiLSTM + multi-level attention [11] | -      | -      | 68.00    | 54.40    |
| Ours                      | 13.2   | 74.31  | 68.53    | 54.28    |

As shown in Table 2, on the CDTB dataset, the advantages of our model are not obvious. When dealing with fine-grained and hierarchical classification, the tree structure can calculate the degree of the matching between the text features and classes layer by layer. However, when there are only four classes, the tree structure does not work, and the improvement of the results is due to the introduction of BERT.

B. FURTHER IDENTIFICATION

To further prove the superiority of our architecture, we removed several key structures from our model and verified their influence on the final results. The results of 4, 19, and 38 classes are reported in Table 3, and the models are as follows: W+B+T+I (the same architecture as our method but it uses the Word2Vec instead of BERT), B+R (BERT and RNN), B+B+S (the same architecture as our method but it applies the softmax layer instead of tree structure and improved loss function), B+B+T+C (the same architecture as our method but it uses the cross entropy loss function), and ours.

The results of different structures show that every part of our model is necessary for the entire architecture. Compared with Word2Vec, the BERT word embedding extracts discourse features before the classification model. Our method achieves results that are better than those of the model without a tree structure (B+B+S) because the classification results of each level interact with each other owing to the loss function in the tree structure. The results of B+R and B+B+S demonstrate that BiLSTM is better than the RNN when integrated with BERT and applied to the task of long text semantic analysis. Moreover, our improved loss function is better than the cross entropy loss in the tree structure.

We showed the F1 score of each compared model in Figure 5. From Table 3 and Figure 5, the larger the number of classes is, the more obvious the effects of BERT and tree structure. The phenomenon mainly stems from the fact that the data imbalance is obvious when the number of class is large.

C. DISCUSSION

Compared with an RNN or RNN with attention mechanism, the tree structure and improved loss function of our model show their superiority in the task for Chinese implicit DRR on the HIT-CDTB. This illustrates that the relation-
ship tree structure is effective for a fine-grained semantic discourse classification. However, when the classification of the dataset is not hierarchical, the tree structure cannot be generated. Thus, for the 4-class classification task on the CDTB, the improvement of the results is not obvious.

From the results of different model structures, our model can complete the task of discourse semantic analysis in the case of unbalanced data. The pretraining of BERT can obtain as much feature information in a long discourse as possible without considering the amount of data in one class. The tree structure enhances its classification constraints and reduces the negative effects of a data imbalance through the interaction of each class level. Moreover, compared with other models with the similar complexity, our model has higher time efficiency.

Above all, the proposed model can effectively deal with the problem of data imbalance when the classification of the dataset is not hierarchical.

VI. CONCLUSION

In this study, we presented a Chinese implicit DRR model based on BERT word embedding and the classification representation of a tree structure, and solved the problem of data imbalance when the number of classes is large. The experiment results demonstrate that BERT and the proposed tree structure can complete the task for Chinese implicit DRR task effectively. The ability of our method to model discourse tree structure can complete the task for Chinese implicit DRR, and solved the problem of unbalanced data. The pretraining of BERT can obtain as much feature information in a long discourse as possible without considering the amount of data in one class. The tree structure enhances its classification constraints and reduces the negative effects of a data imbalance through the interaction of each class level. Moreover, compared with other models with the similar complexity, our model has higher time efficiency.

In a future study, we intend to extend our approach to different languages and domains. We believe that the classification constraint of the tree structure of the discourse information may be a driving force in successfully handling complex semantic processing tasks.

REFERENCES

[1] M. Sun and J. Y. Chai, “Discourse processing for context question answering based on linguistic knowledge,” Knowl.-Based Syst., vol. 20, no. 6, pp. 511–526, Aug. 2007.
[2] R. S. Trivedi and J. Eisenstein, “Discourse connectors for latent subjectivity in sentiment analysis,” in Proc. NAACL, Atlanta, GA, USA, 2013, pp. 808–813.
[3] M. Tu, Y. Zhou, and C. Zong, “A novel translation framework based on rhetorical structure theory,” in Proc. ACL, Sofia, Bulgaria, 2013, pp. 370–374.
[4] E. Pitter, A. Louis, and A. Nenkova, “Automatic sense prediction for implicit discourse relations in text,” in Proc. ACL AfNLP, Singapore, 2009, pp. 683–691.
[5] R. Elwell and J. Baldridge, “Discourse connective argument identification with connective specific rankers,” in Proc. IEEE Int. Conf. Semantic Comput., Santa Monica, CA, USA, Aug. 2008, pp. 198–205.
[6] D. Marcu and A. Echihabi, “An unsupervised approach to recognizing discourse relations,” in Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL), Philadelphia, PA, USA, 2001, pp. 368–375.
[7] F. Kong and G. Zhou, “A CDT-styled end-to-end Chinese discourse parser,” ACM Trans. Asian Low-Resource Lang. Inf. Process., vol. 16, no. 4, pp. 1–17, Sep. 2017.
[8] S. Römövquist, N. Schenk, and C. Chiarosc, “A recurrent neural model with attention for the recognition of chinese implicit discourse relations,” in Proc. ACL, Vancouver, BC, Canada, 2017, pp. 256–262.
[9] H. Li, J. Zhang, and C. Zong, “Implicit discourse relation recognition for english and chinese with multiview modeling and effective representation learning,” ACM Trans. Asian Low-Resource Lang. Inf. Process., vol. 16, no. 3, pp. 1–21, Apr. 2017.
[10] S. Xu, P. Li, F. Kong, Q. Zhu, and G. Zhou, “Topic tensor network for implicit discourse relation recognition in chinese,” in Proc. ACL, Florence, Italy, 2019, pp. 608–618.
[11] Y. Liu and S. Li, “Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention,” in Proc. Conf. Empirical Methods Natural Lang. Process., Austin, TX, USA, 2016, pp. 1224–1233.
[12] A. Rutherford, V. Dembo, and N. Xue, “A systematic study of neural discourse models for implicit discourse relation,” in Proc. EACL, Valencia, Spain, 2017, pp. 281–291.
[13] R. Prasad, S. Husain, D. Sharma, and J. Aravind, “Towards an annotated corpus of discourse relations in Hindi,” in Proc. IJCNLP, Hyderabad, India, 2008, pp. 73–80.
[14] D. Zeyrek and B. Webber, “A discourse resource for Turkish: Annotating discourse connectives in the METU corpus,” in Proc. IJCNLP, Hyderabad, India, 2008, pp. 65–72.
[15] A. AlSaif and K. Markert, “The leeds Arabic discourse treebank: Annotating discourse connectives for Arabic,” in Proc. LREC, Valletta, Malta, 2010, pp. 2046–2053.
[16] L. Carlson, D. Marcu, and M. E. Okurowski, “Building a discourse-tagged corpus in the framework of rhetorical structure theory,” in Proc. 2nd Sigdial Workshop Discourse Dialogue, 2001, pp. 85–112.
[17] R. Prasad, N. Dinesh, A. Lee, M. Mitsakaki, L. Robaldo, A. Joshi, and B. Webber, “The Penn discourse treebank 2.0,” in Proc. LREC, Marrakech, Morocco, 2008, pp. 2961–2968.
[18] Y. Zhou and N. Xue, “PDDB-style discourse annotation of Chinese text,” in Proc. ACL, Jeju, South Korea, 2012, pp. 69–77.
[19] Y. Li, W. Feng, J. Sun, F. Kong, and G. Zhou, “Building chinese discourse corpus with connective-driven dependency tree structure,” in Proc. EMNLP, Doha, Qatar, 2014, pp. 2105–2114.
[20] M. Zhang, B. Qin, and T. Liu, “Chinese discourse relation semantic taxonomy and annotation,” in (Chinese), J. Chin. Inf. Process., vol. 28, no. 2, pp. 28–36, 2014.
[21] A. Rutherford and N. Xue, “Robust non-explicit neural discourse parser in English and Chinese,” in Proc. ACL, Berlin, Germany, 2016, pp. 55–59.
[22] Y. Xu and Y. Wu, “Chinese discourse relation recognition using parallel corpus,” in Proc. 9th Int. Conf. Comput. Intell. Secur., Sichuan, China, Dec. 2013, pp. 424–427.
[23] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL, Minneapolis, MN, USA, 2019, pp. 4171–4186.
[24] L. Lee, Y. Lu, P. Chen, P. Lee, and K. Shyu, “NCUERE at MEDIQA 2019: Medical text inference using ensemble BERT-BiLSTM-attention model,” in Proc. BioNLP, Sichuan, China, 2019, pp. 528–532.
[25] Y. Ohsugi, I. Saito, K. Nishida, H. Asano, and J. Tomita, “A simple but effective method to incorporate multi-turn context with BERT for conversational machine comprehension,” in Proc. ACL, Florence, Italy, 2019, pp. 11–17.
[26] B. Didenko and J. Shaptala, “Multi-headed architecture based on BERT for grammatical errors correction,” in Proc. ACL, Florence, Italy, 2019, pp. 246–251.
[27] Z. Gao, A. Feng, X. Song, and X. Wu, “Target-dependent sentiment classification with BERT,” IEEE Access, vol. 7, pp. 154290–154299, 2019.
[28] C. Zhang and M. Abdul-Mageed, “No army, no navy: BERT supervised learning of arabic dialects,” in Proc. ACL, Florence, Italy, 2019, pp. 279–284.
[29] D. Kondratyuk, “Cross-lingual lemmatization and morphology tagging with two-stage multilingual BERT fine-tuning,” in Proc. ACL, Florence, Italy, 2019, pp. 12–18.
[30] L. Yao, Z. Jin, C. Mao, Y. Zhang, and Y. Luo, “Traditional chinese medicine clinical records classification with BERT and domain specific corpora,” J. Amer. Med. Inform. Assoc., vol. 26, no. 12, pp. 1632–1636, Dec. 2019.
[31] D. Zhang, L. Tian, M. Hong, F. Han, Y. Ren, and Y. Chen, “Combining convolution neural network and bidirectional gated recurrent unit for sentence semantic classification,” IEEE Access, vol. 6, pp. 73750–73759, 2018.
[32] V. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal loss for dense object detection,” in Proc. ICCV, Venice, Italy, Oct. 2017, pp. 2999–3007.
DAN JIANG was born in Heilongjiang, China, in 1986. She is currently pursuing the Ph.D. degree with the Beijing University of Posts and Telecommunications. Her research interests include natural language processing and deep learning.

JIN HE was born in Shanxi, China, in 1988. He is currently pursuing the Ph.D. degree with the Beijing University of Posts and Telecommunications. His research interests include artificial intelligence and deep learning.

* * *