Constructing Chinese Macro Discourse Tree via Multiple Views and Word Pair Similarity

Yi Zhou, Xiaomin Chu, Peifeng Li and Qiaoming Zhu

School of Computer Sciences and Technology, Soochow University, Jiangsu, China
{yzhou0928,xmchu}@stu.suda.edu.cn;{pfli,qmzhu}@suda.edu.cn

Abstract. Macro-discourse structure recognition is an important task in macro-discourse analysis. At present, the research on macro-discourse analysis mostly uses the manual features (e.g., the position features), and ignores the semantic information in topic level. In this paper, we first propose a multi-view neural network to construct Chinese macro discourse trees from three views, i.e., the word view, the context view and the topic view. Besides, we propose a novel word-pair similarity mechanism to capture the interaction among the discourse units and the topic. The experimental results on MCDTB, a Chinese discourse corpus, show that our model outperforms the baseline significantly.

Keywords: macro discourse; discourse tree construction; word-pair similarity; multiple views.

1 Introduction

In the field of natural language processing, the granularity of research objects gradually turns to the higher semantic unit, specifically, from the lexical and syntactic analysis on words and sentences to the discourse analysis on sentence groups and paragraphs. The main task of discourse analysis is to clarify the connection between discourse units and explore the logical relationship between them. Discourse analysis is conducive to understanding the organization and the topic of an article, and plays a supporting role for a variety of downstream tasks such as question and answer system [1] and sentiment analysis [2].

There are two levels of discourse analysis on the different granularity of discourse unit. One is the micro discourse analysis, which researches on the relationship among clauses, sentences, and sentence groups, and the other is the macro discourse analysis, which focuses on the relationship between paragraphs and paragraph groups. Macro discourse structure analysis is an important sub-task of macro discourse analysis. In a well-written article, a paragraph should not be isolated but rather organized in a coherent way depend on the context. Based on this fact, referring to the Rhetorical Structure Theory (RST), Chu et al. [3] proposed a framework of macro discourse structure representation in Chinese. It uses a paragraph as an Elementary Discourse Unit (EDU), and these discourse units are merging with their adjacent discourse units to form a new discourse unit. The whole article can be represented as a discourse tree with EDUs as

the leaf nodes. To introduce the representation of the macro discourse tree more intuitively, take chtb_0131 in CTB as an example. (The details of the example are provided in Appendix A)

The discourse structure tree of chtb_0131 is shown in Fig. 1. In this article, the first paragraph presents the main event, and the second and third paragraphs provide data support for it from two different aspects. The last paragraph describes the impact of the main event on other events. In this paper, we mainly explored the macro discourse structure, which is reflected the connection relationship between nodes in Fig. 1.

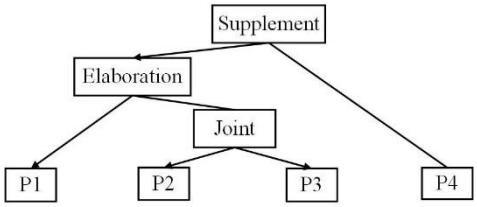

Fig. 1. Macro-discourse tree of chtb_0131.

There are only a few studies [4, 5, 17, 18 19, 25] on macro discourse analysis, and the existing researches have the following two issues. First, they mainly focused on the analysis between the paired of adjacent discourse units, and there are few attempts to construct the overall structure of the macro discourse tree. However, the existed researches have proved that the information supplied by the whole tree structure plays an important role in the nuclearity identification and relationship classification, which are the other two major tasks of macro discourse analysis. Second, most of their semantic information relied on manual features by calculating the similarity of two discourse units, and most of these similarity methods simply averaged word embedding or calculate word similarity in the paragraph as the representation of two paragraphs [4, 5]. These methods failed to consider the coherence of the discourse and may dilute the useful information by forcibly blending all the word information, because the macro discourse unit is longer and contains a great deal of noise information.

To address the above two issues, we propose a multi-view neural network to construct Chinese macro discourse trees. In particular, we introduce three different views, i.e., the word view, the context view and the topic view, to capture the different discourse semantics. Besides, we also propose a novel word-pair similarity mechanism to capture the interaction among the discourse units and the topic. The experimental results on MCDTB, a Chinese discourse corpus, show that our model outperforms the baseline significantly.

2 Related Work

The Rhetorical Structure Theory Discourse Treebank (RST-DT) [6] and the Chinese Macro Discourse Treebank (MCDTB) [17, 18] are two popular corpora for the task of macro discourse analysis.

Based on the Rhetorical Structure Theory (RST) [7, 8], RST-DT annotated 385 articles of the Wall Street Journal selected from the Penn Treebank (PTB) [9]. The research of the discourse structure recognition on this corpus has three levels: intra-sentence, inter-sentence, and inter-paragraph (i.e., macro-level). Hernault et al. [10] proposed a HILDA parser, which used the Support Vector Machine (SVM) to identify discourse units and nuclear-relations, respectively. It is a bottom-up framework of constructing a discourse tree. Feng et al. [11] achieved an excellent performance of identification discourse structure by using two Conditional Random Field (CRF) models with sliding windows. Recently, Morey [12] proposed a method to transform the RST component discourse tree into the dependent discourse tree, which opened up another perspective for discourse tree construction. The neural network models were also used in discourse tree construction. However, most of them focused on micro-level. Li [13] proposed a hierarchical BiLSTM model with the attention mechanism for discourse tree construction, which used a tensor-based transformation method to capture the semantics among discourse units. Jia et al. [14] introduces a memory network into the traditional BiLSTM to capture the topic information of the article. So far, the performances of those neural network methods are still lower than those of the traditional models under the unified evaluation criteria proposed by Morey [15].

There is only one work on macro discourse analysis. Sporleder et al. [16] transforms the RST-DT's discourse trees into the paragraph-level macro discourse trees, and used the maximum entropy model to build discourse trees.

MCDTB [17, 18] is a Chinese macro discourse corpus, annotating the structure, nuclearity, and relationship of macro discourse structure. Currently, MCDTB contains 720 news documents annotated with 3 categories (remove transition for adapting to the macro discourse structure) and 15 relations. Jiang et al. [5] proposed a CRF-based joint model for the structure recognition and nuclear identification of macro discourses. The experimental results showed the importance of discourse position information and the sub-tree structure information in the task of judging the relationship between a pair of discourse units. Chu et al. [19] used the Integer Linear Programming (ILP) to coordinate the relationship between nuclear and structure. Specifically, they trained two CRF models for structure recognition and nuclear identification, respectively. However, they only recognize the macro structure between two or more discourse units and did not construct a complete macro discourse tree.

3 Multi-view Model on Word-pair Similarity

In this paper, we employ the popular transition based approach (shift-reduce) [20] to construct macro discourse tree. In a typical shift-reduce approach for discourse parsing, the parsing process is modeled as a sequence of *shift* and *reduce* actions on a stack and a queue. The shift-reduce approach is to determine whether a discourse unit is more likely to merge with its previous discourse unit or following discourse unit. Following previous approaches, we select the top two discourse units (i.e., S1 and S2) in the stack and the first discourse unit (i.e., Q1) in the queue as the input of our model.

Discourse semantics is the core evidence to judge whether merging two discourse

units. In this paper, we introduce three views, the word view, the context view and the topic view, to represent the discourse semantics. Basically, the semantics of a discourse unit originates from its containing words. Hence, we introduce the word view to our model to represent the semantics of isolate words in the discourse units. Moreover, to understand the meanings of an article, Humans maybe need to read it many times because its meaning not only derives from the isolate words, but also depends on their contexts. In our model, the representation of the hidden layer of LSTM in each time step can be regarded as a context view of word semantics. However, whether merging two discourse units is not only related to the semantics of the discourse units, but also related to the relationship between the topic of the entire article and a discourse unit. Hence, we also introduce the title of the article to represent the semantic view of the topic, i.e., the topic view, as the additional input.

The structure of our multi-view model on word-pair similarity is shown in Fig. 2, including three parts: a text encoding network, a word-pair similarity mechanism, and a binary action classifier.

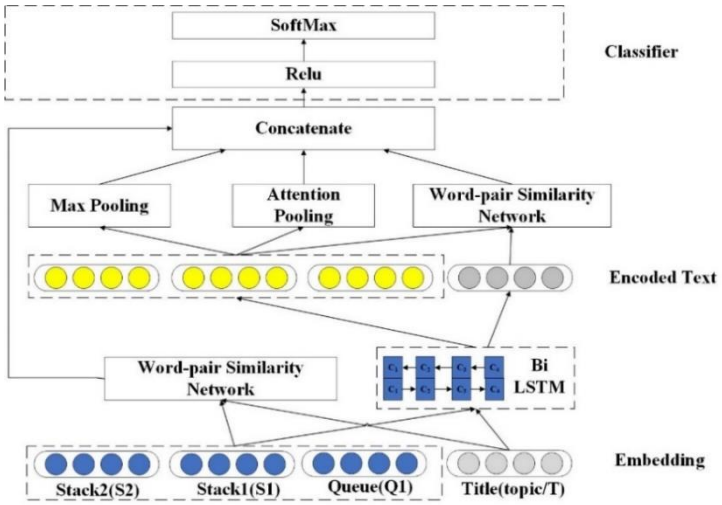

Fig. 2. The structure of our multi-view model on word-pair similarity.

3.1 Shift-reduce Algorithm for Discourse Tree Construction

Shift-reduce approach transforms the procedure of tree construction into a sequence of two actions, *shift* and *reduce*. It uses a queue and a stack. First, puts all EDUs into the queue. At each step, it performs one of the *shift* or *reduce* actions. The *shift* action pushes the first unit in the queue into the stack, and the *reduce* action merges the top two units in the stack into a larger unit and then pushes the merged unit back to the top of the stack. Repeat the step until the queue is empty and the stack contains only one unit, and the only unit in the stack is the root node of the whole tree. At each step, it employs our multi-view model to select the actions.

3.2 Text Encoding Network

The input of our multi-view model is the word sequences of S1, S2, Q1 and the title topic (T) and can be represented as $s=(s_1, s_2, \ldots, s_N)$ where N is the number of the words in a word sequence. Firstly, we pre-train the word embeddings with Word2Vec on the Wikipedia Chinese corpus and convert the four word sequences to four word vectors where $\boldsymbol{D}_{word}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_N)$ can be regarded as the word view ($\boldsymbol{D}\in\{\boldsymbol{S1}, \boldsymbol{S2}, \boldsymbol{Q1}\}$) and the topic view ($\boldsymbol{D}=\boldsymbol{T}$).

Since the macro discourse unit is long, a localized model such as CNN may suffer from the redundant information. In those macro discourses, the interaction with the adjacent discourses occurs more at the beginning and end of the discourse, so the Bidirectional LSTM (BiLSTM) is used as the encoding layer because it pays more attention to the words at the beginning and the end of the sequences.

The input of BiLSTM is the word vector $(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_N)$ of the discourse unit (S1, S2 and Q1) or title (T). At each timestep, the results of the two LSTMs $\overrightarrow{\boldsymbol{h}_f},\overleftarrow{\boldsymbol{h}_b}\in\mathbb{R}^l$ are concatenated as follows to obtain a context-dependent representation of a word $\boldsymbol{w}_i^{\,h}\in\mathbb{R}^{2l}$.

$$\boldsymbol{w}_i^{\,h}=[\overrightarrow{\boldsymbol{h}_f},\overleftarrow{\boldsymbol{h}_b}] \tag{1}$$

where l is the number of hidden layer units in LSTM, and $\boldsymbol{D}_{context}=(\boldsymbol{w}_1^{\,h},\boldsymbol{w}_2^{\,h},\cdots,\boldsymbol{w}_N^{\,h})$ represents the context view ($\boldsymbol{D}\in\{\boldsymbol{S1}, \boldsymbol{S2}, \boldsymbol{Q1}\}$) and the topic view ($\boldsymbol{D}=\boldsymbol{T}$), respectively.

Finally, the maximum pooling and attention pooling are performed on the hidden layer states of all timesteps $\boldsymbol{H}$ to get the discourse representation $\boldsymbol{v}_{max}$ and $\boldsymbol{v}_{att}$, and the two results are concatenated as the representation of the discourse unit vector $\boldsymbol{v}\in\mathbb{R}^{4l}$ as follows.

$$\boldsymbol{v}_{max}=maxpooling(\boldsymbol{H}) \tag{2}$$

$$\boldsymbol{v}_{att}=sum(softmax(\boldsymbol{H}\boldsymbol{W}_{att}+\boldsymbol{b}_{att})\otimes\boldsymbol{H}) \tag{3}$$

$$\boldsymbol{v}=[\boldsymbol{v}_{max},\boldsymbol{v}_{att}] \tag{4}$$

where the $\otimes$ operation represents the element-wise multiplication, $\boldsymbol{W}_{att}$ and $\boldsymbol{b}_{att}$ are parameters of the attention layer, and sum represents the operation summing the results of each timestep by each dimension.

3.3 Word-pair Similarity Mechanism

After inputting the word sequences into the text encoding network, we obtain the representations of discourse unit from three views, i.e., the representation of the word view $\boldsymbol{D}_{word}$ ($\boldsymbol{D}\in\{\boldsymbol{S1}, \boldsymbol{S2}, \boldsymbol{Q1}\}$), the representation of the context view $\boldsymbol{D}_{context}$ ($\boldsymbol{D}\in\{\boldsymbol{S1}, \boldsymbol{S2}, \boldsymbol{Q1}\}$) and the representation of the topic view $\boldsymbol{D}_{word}$ ($\boldsymbol{D}=\boldsymbol{T}$) and $\boldsymbol{D}_{context}$ ($\boldsymbol{D}=\boldsymbol{T}$), as follows.

$$\boldsymbol{D}_{word}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_N) \tag{5}$$

$$\boldsymbol{D}_{context}=(\boldsymbol{w}_1^{\,h},\boldsymbol{w}_2^{\,h},\cdots,\boldsymbol{w}_N^{\,h}) \tag{6}$$

where $\boldsymbol{D}_{word}\in\mathbb{R}^{N\times d}$ and $\boldsymbol{D}_{context}\in\mathbb{R}^{N\times 2l}$, d is the dimension of word embedding.

The studies on micro discourse analysis (e.g., Lin et al. [21]) demonstrated that the word pair features are effective in the traditional machine learning methods. Currently,

the studies on macro discourse analysis focus on the representation of the entire discourse unit and ignore the interactive information on word pairs. Inspired by the studies on discourse relation recognition [22], we proposes a word-pair similarity mechanism to capture the interaction between different discourse units, as shown in Figure 3.

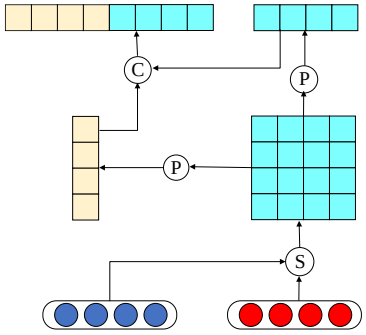

Fig. 3. Word-pair Similarity Network.

The input of the word-pair similarity network is two word matrices $\boldsymbol{DU}_1^{view}$ and $\boldsymbol{DU}_2^{view}$ ($\boldsymbol{DU}_1^{view}, \boldsymbol{DU}_2^{view} \in \boldsymbol{D}_{view}$ where $\boldsymbol{view} \in \{\boldsymbol{word}, \boldsymbol{context}\}$) with the same dimension. It first calculates the similarity of each word in the input word matrices and gets the word pair similarity matrix $\boldsymbol{SimMatrix}^{view} \in \mathbb{R}^{N \times N}$. Then, it pools the similarity matrix in the horizontal and vertical directions to obtain the similarity of each word in one of the discourse units with the other, which can be noted as $\boldsymbol{Simh}^{view}, \boldsymbol{Simv}^{view} \in \mathbb{R}^{N}$ as follows.

$$\boldsymbol{Simh}^{view} = PoolingH(sim(\boldsymbol{DU}_1^{view}, \boldsymbol{DU}_2^{view})) \tag{7}$$

$$\boldsymbol{Simv}^{view} = PoolingV(sim(\boldsymbol{DU}_1^{view}, \boldsymbol{DU}_2^{view})) \tag{8}$$

where *PoolingH* and *poolingV* represent the performing pooling function in the horizontal and vertical direction, respectively. *sim* represents the similarity calculation in the word-pair similarity network. We choose the same maximum pooling and attention pooling mechanism as Subsection 3.2.

Following Xu et al [23] on micro-discourse analysis, we choose the cosine distance and bilinear as the similarity calculation function *sim* as follows.

$$cosine(\boldsymbol{DU}_1^{view}, \boldsymbol{DU}_2^{view}) = \frac{\boldsymbol{DU}_1^{view} \cdot \boldsymbol{DU}_2^{view}}{\left|\boldsymbol{DU}_1^{view}\right|\left|\boldsymbol{DU}_2^{view}\right|} \tag{9}$$

$$bilinear(\boldsymbol{DU}_1^{view}, \boldsymbol{DU}_2^{view}) = (\boldsymbol{DU}_1^{view})^T \boldsymbol{W} (\boldsymbol{DU}_2^{view}) \tag{10}$$

where $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ ($\boldsymbol{view}$=$\boldsymbol{word}$) or $\boldsymbol{W} \in \mathbb{R}^{2l \times 2l}$ ($\boldsymbol{view}$=$\boldsymbol{context}$) is a parameter matrix with random initialization.

Finally, it concatenates them to obtain the final discourse similarity representation of $\boldsymbol{DU}_1$ and $\boldsymbol{DU}_2$ under the view $\boldsymbol{view}$ as follows.

$$\boldsymbol{SimVec}_{DU_1, DU_2}^{view} = [\boldsymbol{Simv}^{view}, \boldsymbol{Simh}^{view}] \tag{11}$$

3.4 Action Classifier

Using word-pair similarity network, we can obtain the similarity from the word view ($\boldsymbol{SimView}_{du}^{word}$), the context view ($\boldsymbol{SimView}_{du}^{context}$) and the topic view ($\boldsymbol{SimView}_{topic}^{word}$, $\boldsymbol{SimView}_{topic}^{context}$) as follows.

$$\boldsymbol{SimView}_{du}^{word} = [\boldsymbol{SimVec}_{S2,S1}^{word}, \boldsymbol{SimVec}_{S1,Q1}^{word}, \boldsymbol{SimVec}_{S2,Q1}^{word}] \tag{12}$$

$$\boldsymbol{SimView}_{du}^{context} = [\boldsymbol{SimVec}_{S2,S1}^{context}, \boldsymbol{SimVec}_{S1,Q1}^{context}, \boldsymbol{SimVec}_{S2,Q1}^{context}] \tag{13}$$

$$\boldsymbol{SimView}_{topic}^{word} = [\boldsymbol{SimVec}_{S2,topic}^{word}, \boldsymbol{SimVec}_{S1,topic}^{word}, \boldsymbol{SimVec}_{Q1,topic}^{word}] \tag{14}$$

$$\boldsymbol{SimView}_{topic}^{context} = [\boldsymbol{SimVec}_{S2,topic}^{context}, \boldsymbol{SimVec}_{S1,topic}^{context}, \boldsymbol{SimVec}_{Q1,topic}^{context}] \tag{15}$$

Then we concatenate the representations of the above three views and the representations of three isolate discourse units $\boldsymbol{v}_{S2}, \boldsymbol{v}_{S1}, \boldsymbol{v}_{Q1}$ together to form the feature vector $\tilde{\boldsymbol{v}}$ as follows.

$$\boldsymbol{D} = [\boldsymbol{v}_{S2}, \boldsymbol{v}_{S1}, \boldsymbol{v}_{Q1}] \tag{16}$$

$$\tilde{\boldsymbol{v}} = [\boldsymbol{D}, \boldsymbol{SimView}_{du}^{word}, \boldsymbol{SimView}_{topic}^{word}, \boldsymbol{SimView}_{du}^{context}, \boldsymbol{SimView}_{topic}^{context}] \tag{17}$$

Finally, the result is obtained by applying a binary classifier with Relu Layer on features as follows.

$$\boldsymbol{t} = Relu(\tilde{\boldsymbol{v}}\boldsymbol{W}_r + \boldsymbol{b}_r) \tag{18}$$

$$\boldsymbol{pred} = Softmax(\boldsymbol{t}\boldsymbol{W}_p + \boldsymbol{b}_p) \tag{19}$$

where $\boldsymbol{W}_r \in \mathbb{R}^{(6l+18N+6N_t)\times hdim}$, $\boldsymbol{b}_r \in \mathbb{R}^{hdim}$, $\boldsymbol{W}_r \in \mathbb{R}^{hdim\times 2}$, $\boldsymbol{b}_p \in \mathbb{R}^2$ are parameter matrices, N_t represents the number of words contained in the topic, and *hdim* represents the number of hidden units of the fully connected layer.

4 Experiments

In this section, we first introduce the experimental dataset and setting, and then report the experimental results and gives the analysis.

4.1 Dataset and Experimental Setting

We conducted our experiments on the Macro Chinese Discourse Treebank (MCDTB). This corpus annotated 720 articles from CTB 8.0, including a total of 3,981 paragraphs, 8,319 sentences, and 398,829 words. The paragraph lengths of the articles are from 2 to 22, as showed in Table 1.

Table 1. Distribution of article length in MCDTB (in paragraph).

Length	2	3	4	5	6	7	8	9	10	11	12	>13
Number	29	112	159	144	91	58	37	33	15	13	14	15

Following Chu [19], we divide the data into 10 sets to achieve the balance of length distribution on each set and use the ten-fold cross validation in our evaluation. In each folder, the data split is 8:1:1 for training, validation and test.

In our experiments, we use the standard word segmentation results annotated by CTB8.0. Since the shift-reduce approach finally generates a binary tree, we convert the multi-fork trees into the right-heavy binary trees, following the related work on RST-DT [11, 15, 24]. Fig. 4 is an example to convert a multi-fork tree (left) to a binary trees (right).

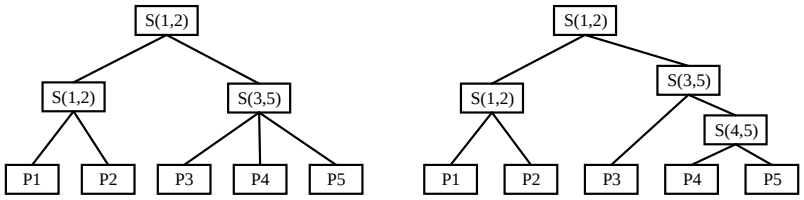

Fig. 4. The right-heavy binarization of macro-discourse tree.

We pre-trained the word embeddings with Word2Vec on the Wikipedia Chinese corpus and set 50 to the dimension. According to the experiments conducted on the development set, the number of hidden layer units in BiLSTM is set to 50, and the number of hidden layer units of the Relu layer is determined as follows:

$$hiddenUnitNum = \max(1024, \lfloor featureSize / 10 \rfloor) \tag{20}$$

The minibatch approach is used in our training and the batch size is set to 96 and the training epoch is set to 30. Finally, the model with the best accuracy on the validation dataset is selected to evaluate the test dataset.

We used the right-heavy binary tree as the gold data for evaluation. Following Morey et al. [15], we use internal node accuracy (equal to micro-F1) as the evaluation metric objective, which evaluates how likely discourse units are correctly merged.

4.2 Experimental Result

Because the existing work on macro discourse tree construction only judged whether there was a relationship between two completely correct DUs, they cannot be directly used as baseline. Hence, we reproduced Jiang's degradation model [5]. The experimental results are shown in Table 2, where the MVM is our multi-view model.

Table 2. The performance comparison on MCTDB.

Name	NodeAcc (%)
Jiang	54.21
MVM(cosine)	**58.77**
MVM(bilinear)	56.12

Table 2 shows that our model MVM outperforms the baseline Jiang on the internal

node accuracy by 4.56, and this result ensures that our multi-view neural network model can capture the discourse semantics from three layers, i.e., the word, context and topic, to improve the performance of discourse tree construction. Compared with Jiang using manual methods to extract similarity features, our MVM only uses the simple discourse units and the topic as input. This also proves the feasibility and effectiveness of the neural network model to construct discourse trees. It should also be noted that bilinear similarity shows worse performance than simple cosine similarity. It is because the scale of the corpus is too small to learn such a large number of additional parameters.

4.3 Analysis

To explore the effectiveness of different views, Table 3 shows the comparison of different simplified models. From Table 3, we can find out that the word view, the context view and the topic view can improve the internal node accuracy simultaneously. This result ensures that all of three views are helpful for discourse tree construction.

Table 3. The comparison of different simplified models with MVM.

Name	Description	NodeAcc (%)
Baseline	Removing all three views from MVM	54.09
Baseline + word view	Adding the word view to the baseline	55.19
Baseline + context view	Adding the context view to the baseline	57.15
Baseline + topic view	Adding the topic view to the baseline	56.57
MVM (Baseline + all views)	Our multi-view model	**58.77**

Table 3 shows that the improvement of the word view is lower than that of the context view (1.1 vs 3.06). This result can conclude that the overall semantic tendency is more important than independent vocabulary in macro-structure identification. Table 3 also shows that the topic view also improves the internal node accuracy by 2.48 and this result shows that the relationship between the topic and the discourse units is also helpful for macro-discourse structure recognition.

To verify the effectiveness of our word-pair similarity mechanism, we compare it with two other discourse similarity mechanisms, Cosine distance and the mechanism used in Jiang et al. [25]. The Cosine distance calculates the angle between two vectors, which is usually used to measure the degree of similarity. The similarity in [25] is a method to calculate the similarity of texts based on the word vector. Table 4 shows the results using different similarity mechanism and it shows that our model MVM outperforms the other two mechanisms on the internal node accuracy by 4.41 and 2.2, respectively. This result shows that our word-pair similarity mechanism is better to capture the difference between two discourse units and between the discourse unit and the topic.

Both the cosine distance and Jiang similarity represent the discourse units integrating all words or word pairs in the discourses, and finally it is reduced to a float value. Table 3 shows that their mechanisms are not suitable for macro discourse due to two reasons. The first one is that the amount of information contained in a macro discourse is relative

huge and it is difficult to express the key information by using manual features. The second one is that there is a huge amount of noise information in macro discourse units. Their mechanisms simply fused all the similarities and this will make the noise information pollute the similarity features. On the contrary, our word-pair similarity and neural network model can redistribute the similarity on multiple views and then reduce the influence of noise.

Table 4. The performance comparison with other similarity calculation method.

Name	NodeAcc (%)
MVM with word-pair similarity	**58.77**
MVM with Discourse Cosine	54.36
MVM with Jiang similarity	56.57

To explore the information captured by our word-pair similarity mechanism, we plot the heat map of the similarity matrix on a sample, as shown in Fig. 5. The brightness of the three heat maps in the first row shows the similarity of each word pair between S2-S1, S2-Q1 and S1-Q1 from the left to the right under the word view, and the second line shows the corresponding heat maps under the context view.

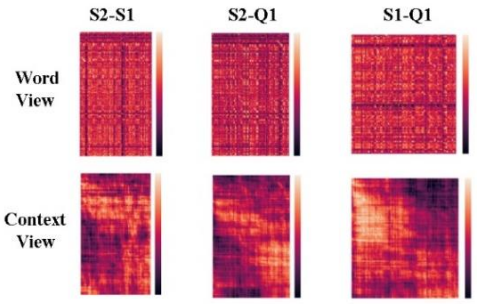

Fig. 5. Heat map of the word-pair-level similarity matrix.

Under the word view, our model is more concerned with the words with strong interaction, which can be visualized as black and white bars in the heat map. In contrast, the semantic transformation becomes softer and shows a more clear light area under the context view. This means that the context view can weaken the ability of indicating the absolute position of the keywords and enhance the ability to express the interaction of two discourse units.

Fig. 5 shows a *shift* action in which S1 and Q1 have the joint relation (The details of the example are provided in Appendix A). This heat map shows that the similar area between S2 and S1 (S2-S1) is concentrated in its upper part. That is, S2 is more relevant with the first half part of S1. Meanwhile, the similar area between S2 and Q1 (S2-Q1) is concentrated in the lower right part of the heat map. It indicates Q1 is more relevant with the second half part of S1. Finally, the similar area between S1 and Q1 is concentrated in the upper left corner. It shows that the beginning of the two discourses are

similar, while their other parts are not similar. Hence, this is a typical joint relationship, where S1 and Q1 describe the two aspects of S2, respectively.

5 Conclusion

In this paper, we propose a multi-view neural network to construct Chinese macro discourse trees. In particular, we introduce three different views, i.e., the word view, the context view and the topic view, to capture the different discourse semantics. Besides, we also propose a novel word-pair similarity mechanism to capture the interaction among the discourse units and the topic. The experimental results on MCDTB, a Chinese discourse corpus, show that our model outperforms the baseline significantly. Our future work will focus on finding a better representation view for a small-scale corpus.

Acknowledgments. The authors would like to thank three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China under Grant Nos. 61836007, 61772354 and 61773276.

References

1. Galitsky, B., Ilvovsky, D.: Building dialogue structure from discourse tree of a question. In: Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI. pp. 17–23 (2018).
2. Kraus, M., Feuerriegel, S.: Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees. Expert Systems with Applications. 118, 65–79 (2019).
3. Chu, X., Research on representation schema, resource construction and computational modeling of macro discourse structure. Doctorate dissertation, Soochow University, Suzhou, (2019). [in Chinese]
4. Zhou, Y., Chu, X., Zhu, Q., Jiang, F., Li, P. Macro discourse-level relation classification based on macro semantics representation. Journal of Chinese Information Processing, 33:1-7+24(2019). [in Chinese]
5. Jiang, F., Li, P., Chu, X., Zhu, Q., Zhou, G.: Recognizing macro Chinese discourse structure on label degeneracy combination model. In: CCF International Conference on Natural Language Processing and Chinese Computing. pp. 92–104 (2018).
6. Carlson, L., Marcu, D., Okurowski, M.: RST discourse treebank. Linguistic Data Consortium. (2002).
7. Mann, W.C., Thompson, S.A.: Relational propositions in discourse. Discourse Processes. 9, 57–90 (1986).
8. Mann, W.C., Thompson, S.A.: Rhetorical structure theory: Toward a functional theory of text organization. Text-Interdisciplinary Journal for the Study of Discourse. 8, 243–281 (1988).
9. Marcus, M., Sanrotini, B., Marcinkiewicz, M.: Building a large annotated corpus of English: the Penn Treebank. Computational Linguistics. 19, 313–330 (1993).
10. Hernault, H., Prendinger, H., Ishizuka, M.: HILDA: A discourse parser using support vector machine classification. Dialogue and Discourse. 1, 1–33 (2010).

11. Feng, V.W., Hirst, G.: A linear-time bottom-up discourse parser with constraints and post-editing. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 511–521 (2014).
12. Morey, M., Muller, P., Asher, N.: A dependency perspective on RST discourse parsing and evaluation. Computational Linguistics. 44, 197–235 (2018).
13. Li, Q., Li, T., Chang, B.: Discourse parsing with attention-based hierarchical neural networks. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 362–371 (2016).
14. Jia, Y., Ye, Y., Feng, Y., Lai, Y., Yan, R., Zhao, D.: Modeling discourse cohesion for discourse parsing via memory network. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 438–443 (2018).
15. Morey, M., Muller, P., Asher, N.: How much progress have we made on RST discourse parsing? A replication study of recent results on the RST-DT. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 1319–1324 (2017).
16. Sporleder, C., Lascarides, A.: Combining hierarchical clustering and machine learning to predict high-level discourse structure. In: Proceedings of the 20th International Conference on Computational Linguistics. (2004).
17. Chu, X., Jiang, F., Xu, S., Zhu, Q.: Building a macro Chinese discourse treebank. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (2018).
18. Jiang, F., Xu, S., Chu, X., Li, P., Zhu, Q., Zhou, G.: MCDTB: A macro-level Chinese discourse treebank. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 3493–3504 (2018).
19. Chu, X., Jiang, F., Zhou, Y., Zhou, G., Zhu, Q.: Joint modeling of structure identification and nuclearity recognition in Macro Chinese Discourse TreeBank. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 536–546 (2018).
20. Marcu, D.: A decision-based approach to rhetorical parsing. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. pp. 365–372. (1999).
21. Lin, Z., Kan, M.-Y., Ng, H.T.: Recognizing implicit discourse relations in the Penn Discourse Treebank. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1. pp. 343–351. (2009).
22. Guo, F., He, R., Jin, D., Dang, J., Wang, L., Li, X.: Implicit discourse relation recognition using neural tensor network with interactive attention and sparse learning. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 547–558 (2018).
23. Xu, S., Li, P., Zhou, G., Zhu, Q.: Employing text matching network to recognize nuclearity in Chinese discourse. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 525–535 (2018).
24. Joty, S., Carenini, G., Ng, R., Mehdad, Y.: Combining intra-and multi-sentential rhetorical parsing for document-level discourse analysis. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 486–496 (2013).
25. Jiang, F., Chu, X., Xu, S., Li, P., Zhu, Q.: A macro discourse primary and secondary relation recognition method based on topic similarity. Journal of Chinese Information Processing, 32: 43-50 (2018). [in Chinese]