Multi-Grain Representation Learning for Implicit Discourse Relation Recognition

Yu Sun, Huibin Ruan, Yu Hong(✉), Chenghao Wu, Min Zhang, and Guodong Zhou

School of Computer Science and Technology of Soochow University, Suzhou 215006, Jiangsu, China
{sunyu41679, tianxianer, wch759262631, gdzhouatsuda}@gmail.com, huibinruan@outlook.com, zhangminmt@hotmail.com

Abstract. We analyze narrative text spans (also named as arguments) in this paper, and merely concentrate on the recognition of semantic relations between them. Because larger-grain linguistic units (such as phrase, chunk) are inherently cohesive in semantics, they generally contribute more than words in the representation of sentence-level text spans. On the basis of it, we propose the multi-grain representation learning method, which uses different convolution filters to form larger-grain linguistic units. Methodologically, Bi-LSTM based attention mechanism is used to strengthen suitable-grain representation, which is concatenated with word-level representation to form multi-grain representation. In addition, we employ bidirectional interactive attention mechanism to focus on the key information in the arguments. Experimental results on the Penn Discourse TreeBank show that the proposed method is effective.

Keywords: Implicit discourse relation recognition · Multi-grain linguistic units · Bidirectional interactive attention mechanism.

1 Introduction

Implicit discourse relation recognition is a foundational task of Natural Language Processing (NLP), which aims to jointly infer semantic connectives and logical relations between adjacent text spans (also named as arguments) according to semantic information, syntactic information, related domain knowledge and other clues. Implicit discourse relation recognition is helpful for many downstream NLP applications, e.g., question answering [8], machine translation [18], sentiment analysis [20], information extraction [3], etc.

Penn Discourse TreeBank (PDTB) 2.0 [11] is a benchmark corpus for discourse relation recognition. It is mainly defined as four top classes, including Comparison, Contingency, Expansion and Temporal. Previous research mainly used linguistic features and supervised learning methods [7], and word pair made great contributions in their work. Considering the example 1), we naturally infer that the relation between the argument pair is Comparison by word pair (*rose, declined*). However, word pair in some texts is relatively one-sided, as shown in

example 2). The key word pairs are (*good, wrong*) and (*good, ruined*), and the model may simply infer the relation type as Comparison. In fact, "*not a good*" in Arg1 and "*wrong*", "*ruined*" in Arg2 are composed as correct pairs, and then we can correctly infer the relation as Cause based on it. Taking example 3) into account, "*not that significant*" in Arg2 means a little significant instead of slight, the word "*not*" modifies "*that significant*" rather than "*that*" or "*significant*". Thus, it is useful to deal with "*not that significant*" as a whole. On the basis, "*no effect*" and "*not that significant*" are composed as a pair for relation inferring. In short, some larger-grain linguistic units contribute more to the representation of an argument than words. Larger-grain linguistic units are combined with words into multi-grain linguistic units. The units may contain richer semantic information for the task of implicit discourse relation recognition.

1) [*Manufacturers' backlogs of unfilled orders* **rose** *0.5% in September to $497.34 billion*]$_{Arg1}$ [*Implicit=but*] [*Excluding these orders, backlogs* **declined** *0.3%.*]$_{Arg2}$
Relation Type: *Comparison*

2) [*Psyllium's* **not a good** *crop*]$_{Arg1}$ [*Implicit=because*] [*You get a rain at the* **wrong** *time and the crop is* **ruined**.]$_{Arg2}$
Relation Type: *Contingency.Cause*

3) [*The $40 million will have* **no effect** *whatsoever on the asset structure of Eastern's plan*]$_{Arg1}$ [*Implicit=because*] [*Forty million in the total scheme of things is* **not that significant**.]$_{Arg2}$
Relation Type: *Contingency.Cause.Reason*

In this paper, we propose a method of multi-grain representation learning for implicit discourse relation recognition. Convolutional operation can aggregate information of words in a convolutional window. Thus, our method utilizes different convolution filters to form larger-grain linguistic units of an argument. Bi-LSTM based attention mechanism is used to strengthen suitable grained representation which adjusts attention scores of current moment based on the states of the previous moment. We finally obtain arguments represented by different grained linguistic units. And then words are concatenated with them into multi-grain representation which contains richer information. In addition, we introduce the variant of bidirectional attention flow model (BiDAF) [17, 19], an interactive attention mechanism in the field of reading comprehension, into our field as argument interaction.

The rest of the paper is organized as follows. Section 2 summarily concludes related work. Section 3 introduces our approach in detail. Section 4 presents the experimental settings and result analysis. Section 5 concludes the paper.

2 Related Work

PDTB 2.0 which was released by Linguistic Data Consortim (LDC) in February 2008, is a large-scale annotated discourse relation corpus. Since the publication

of the corpus, many researchers [7, 10] have achieved great results based on the linguistic features and supervised learning methods. In recent years, methods based on neural network [2] have achieved significant results in the NLP field.

2.1 Argument Representation

The foundation of the excellent model is representing arguments by an appropriate way. Rutherford et al. [15] used Recursive Neural Network (RNN) to encode context information. Lei et al. [6] combined topic continuity, semantic interaction and attribution to enrich argument representation.

2.2 Argument Interaction

Argument interaction aims to obtain more semantic information, or enhance the key information which can help relation classification between argument pairs. Qin et al. [13] extracted features of the argument pairs through Convolutional Neural Network (CNN), and introduced stacking gated neural architecture to control argument interaction. Lei et al. [6] calculated relation scores between the i-th word of Arg1 and the j-th word of Arg2 respectively over word embedding, and obtained the interaction matrix which represented the relevance of corresponding words in an argument pair. Chen et al. [2] utilized Gated Relevance Network (GRN) to learn interaction between the argument pairs.

Attention mechanism has been widely used in NLP tasks recently. Zhou et al. [21] proposed attention-based Bi-LSTM. Attention mechanism is a method that imitates human reading habit of selectively focusing on partial information. Liu et al. [9] held the idea that humans were unable to focus on important information while read articles at once, thus they grasping the key information of the article required repeated reading and dynamic attention for deciding which was more important at next time. Liu et al. [9] proposed a model of multi-attention mechanism, which achieved the state-of-the-art performance in terms of Temporal and Expansion. Guo et al. [4] proposed interactive attention mechanism.

3 Model

3.1 Overview

The overall architecture of our model is shown in Fig.1, which mainly consists of three parts: word-level layer, larger-grain linguistic units layer and interactive attention layer. In the word-level layer, we take each token of one argument as the input sequence and feed the word of Arg1 and Arg2 to the Bi-LSTM layer. The larger-grain linguistic units layer receives word as input. Firstly, the larger-grain linguistic units are obtained by the convolutional operation with k filters, forming k representations for each argument. Secondly, we utilize the Bi-LSTM based attention mechanism to assign different weights to the k representation of the argument. Furthermore, we sum the k representation up as the final

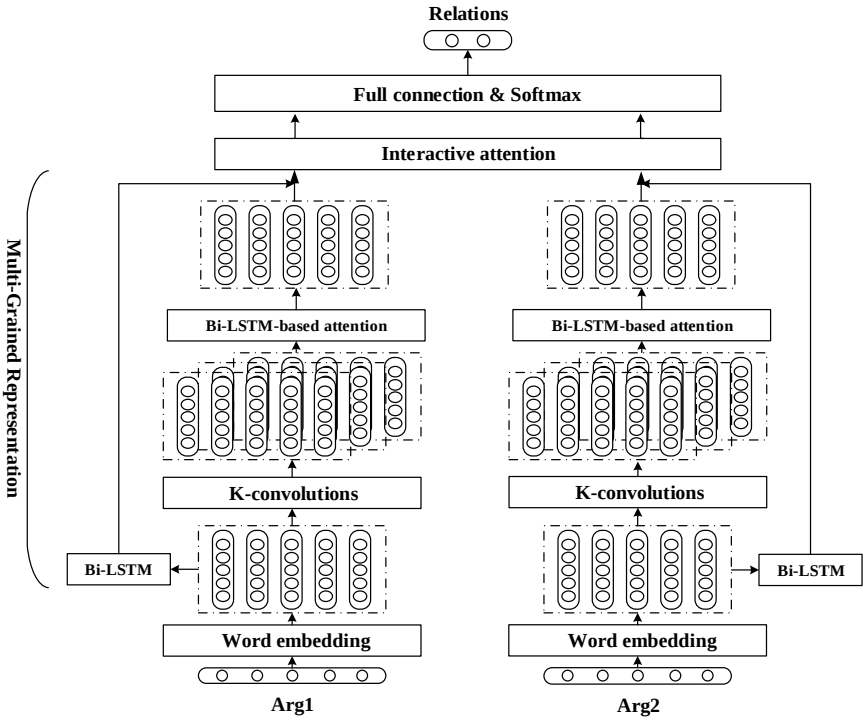

Fig. 1. The overall structure of the model.

representation of the argument. In the interactive attention layer, we concatenate word-level representation and the larger-grain representation as the multi-grain representation, and set it as the input of interactive attention layer. We utilize bidirectional interactive attention mechanism to determine which unit of an argument should be focused on by the information of the other argument. A Bi-LSTM layer and a softmax layer are followed up for the final classification.

3.2 Word Embedding

At the beginning, the words of Arg1 and Arg2 are encoded as fixed-dimensional real-valued vectors by looking up pre-trained word embedding table. Let $x_i^1(x_i^2)$ be the i-th word vector in Arg1(Arg2).

$$X_{Arg1} = [x_1^1, x_2^1, ..., x_s^1] \tag{1}$$

$$X_{Arg2} = [x_1^2, x_2^2, ..., x_s^2] \tag{2}$$

where s denotes the length of an argument, which is fixed and the same for Arg1 and Arg2.

3.3 Word-Level Representation

We set the word embedding representation of the argument as the input of the Bi-LSTM, and obtain the hidden representation at each time. Then, we concatenate these hidden states as the final word-level representation.

$$X'_{Arg1} = BiLSTM(h^1, X_{Arg1}, \theta_1) \tag{3}$$

$$X'_{Arg2} = BiLSTM(h^2, X_{Arg2}, \theta_2) \tag{4}$$

where h^1, h^2 are hidden states. θ_1, θ_2 are learnable parameters of Bi-LSTM.

3.4 Larger-Grain Representation

Larger-Grain Linguistic Units We take advantage of multiple convolutions with different convolution filters. The convolution processes of Arg1 and Arg2 are as follows:

$$c_i = f(w \cdot x_{i:i+h-1} + b) \tag{5}$$

$$C_j = [c_1, c_2, ..., c_s] \tag{6}$$

where $x_{i:i+h-1}$ means word $[x_i,x_{i+1},...,x_{i+h-1}]$; h denotes the size of convolutional window, and $f(\cdot)$ means non-linear function; c_i is the i-th result of convolution; s denotes the length of an argument, and $[\cdot]$ means concatenation operation. C_j is a complete convolutional operation by a convolutional size.

We choose k convolution filters to convolute Arg1 and Arg2 respectively, and then concatenate the k convolutional results as follows:

$$C_{Arg1(2)} = [C_1, C_2, ..., C_k] \tag{7}$$

According to the above operations, we obtain k kinds of representations of an argument which are concatenated by k kinds of larger-grain linguistic units. The units are determined by the size of the convolution filters.

Selecting Appropriate Grained Linguistic Units From above sections, we obtain new vectors C_{Arg1} and C_{Arg2} with k kinds of larger-grain linguistic units. In order to select appropriate grained representation, we adopt Bi-LSTM based attention here. At time t, we calculate the attention weights of different granularities according to the previous hidden state h_{t-1} and the current k linguistic units. As shown in Equation (9,10) [1], the more important granularities at current time are given larger weights.

$$\overrightarrow{h}_t^{1(2)} = \overrightarrow{LSTM}(\overrightarrow{h}_{t-1}^{1(2)}, c_t^{1(2)}, \overrightarrow{\theta}_{3(4)}) \tag{8}$$

$$\overrightarrow{a}_t^{1(2)} = w_{1(2)}^T \tanh(w_{c_{1(2)}} c_t^{1(2)} + w_{h_{1(2)}} \overrightarrow{h}_{t-1}^{1(2)}) \tag{9}$$

$$\overrightarrow{a}_t^{1(2)} = softmax(\overrightarrow{a}_t^{1(2)}) \tag{10}$$

where $\overrightarrow{h}_t^{1(2)}$ is the forward hidden state of Arg1 and Arg2, $\overrightarrow{\theta}_{3(4)}$ means the parameters of the Bi-LSTM. $\overrightarrow{a}_t^{1(2)}$ is the attention score for forward direction, and $w_{1(2)}$, $w_{c_{1(2)}}$ and $w_{h_{1(2)}}$ are learnable parameters. C_i^t, in Equation (11) represents the t-th slice of a convolution result. $c_t^{1(2)}$ means concatenation of k larger-grain linguistic units obtained by different convolution filters. Similarly, the way of calculating reverse direction is as same as the forward.

$$c_t^{1(2)} = C_{Arg1(2)}^t = [C_1^t, C_2^t, ..., C_k^t] \tag{11}$$

The new arguments representation are obtained by weighting the attention mechanism. Then we obtain the larger-grain representation of each argument via concatenating the forward attention result with the reverse one.

$$\overrightarrow{\hat{c}}_t^{1(2)} = \sum_1^k \overrightarrow{a}_t^{1(2)} c_t^{1(2)} \tag{12}$$

$$\overrightarrow{P}_{Arg1(2)} = [\overrightarrow{\hat{c}}_1^{1(2)}, \overrightarrow{\hat{c}}_2^{1(2)}, ..., \overrightarrow{\hat{c}}_s^{1(2)}] \tag{13}$$

$$P_{Arg1(2)} = [\overrightarrow{P}_{Arg1(2)}, \overleftarrow{P}_{Arg1(2)}] \tag{14}$$

3.5 Multi-Grain Representation

The words are concatenated with larger-grain linguistic units, forming the enhanced multi-grain representation, which is merged word-level information with larger-grain information.

$$T_{Arg1} = [X'_{Arg1}, P_{Arg1}] \tag{15}$$

$$T_{Arg2} = [X'_{Arg2}, P_{Arg2}] \tag{16}$$

3.6 Bidirectional Interaction Attention Mechanism

The argument representation obtained from section 3.3 to 3.5 only considers the information of a single argument respectively, and ignore the interactive information between the arguments. Thus, we utilize the interactive learning between argument pairs for further learning. Here, we transfer BiDAF [17, 19] in reading comprehension field to the field of discourse relation recognition. In reading comprehension model, BiDAF is a bidirectional attention mechanism: Query-to-Context and Context-to-Query. Because Query and Context play asymmetric roles to the task, the methods of BiDAF calculating weights for Query and Context are different. In the task of implicit discourse relation recognition, Arg1 and Arg2 are symmetric. So we transform BiDAF to bidirectional interactive attention which is suitable for our task. Thus, each argument can obtain the bidirectional interaction information based on the other argument, and give the greater weights to the important compositions of the argument.

$$T'_{Arg1(2)} = T_{Arg1(2)} \otimes w_{1(2)} \tag{17}$$

$$M = T'_{Arg1} \otimes T'_{Arg2} \tag{18}$$

$$M_{att} = T'_{Arg1} + M + T'_{Arg2} \tag{19}$$

$$O_{Arg11(21)} = softmax(M_{att}) \otimes T_{Arg1(2)} \tag{20}$$

$$O_{Arg12(22)} = softmax(max(M_{att})) \otimes T_{Arg1(2)} \tag{21}$$

$$\begin{aligned} A_{Arg1(2)} =& [T_{Arg1(2)}, O_{Arg11(21)}, T_{Arg1(2)} \odot \\ & O_{Arg11(21)}, O_{Arg11(21)} \odot O_{Arg12(22)}] \end{aligned} \tag{22}$$

where w_1 and w_2 are learnable parameters. M_{att} is a interactive matrix between Arg1 and Arg2. O_{Arg11} and O_{Arg21} are the results for the first method of calculating weights (Context-to-Query), and O_{Arg12} and O_{Arg22} are results for the second (Query-to-Context).

The final representations of arguments are obtained through Bi-LSTM (calculation as section 3.3), which are denoted respectively as H_1 and H_2. Concatenating them into H is as the input of a full-connection layer for feature extraction and dimensionality reduction. Finally, we feed the feature vectors to the softmax layer for classification.

3.7 Model Training

For training, the object is the cross-entropy loss with $L2$ regularization as follows:

$$E(\hat{y}, y) = -\sum_{j}^{s} y_j \times \log(Pr(\hat{y}_j)) \tag{23}$$

$$J(\theta) = \frac{1}{m} \sum_{k}^{m} E(\hat{y}^{(k)}, y^{(k)}) \tag{24}$$

where $Pr(y_j)$ means the probability of assigning the instance to label j, $y^{(k)}$ is the gold labels and $\hat{y}^{(k)}$ is the predicted ones.

4 Experiment

4.1 Dataset and Evaluation Metric

In order to verify the effectiveness of our method, we make use of PDTB 2.0, which is divided into three parts from Rutheford [16], including training set (Section 2-20), development set (Section 0-1) and test set (Section 21-22). All instances test four top-level relation types: Comparison (Comp.), Contingency (Cont.), Expansion (Expa.) and Temporal (Temp.).

The instance number of four types in PDTB is unbalanced. We separately train one-vs-other binary classifiers for each of four discourse relations. A feature

of PDTB dataset is that some instances are annotated as more than one discourse type. We deal with the situation by the way that if the classifier predicts the instance as one of the annotated types, then the prediction will be regarded as correct. We adopt F1-scores for model evaluation.

4.2 Parameter Settings

For the hyper-parameters of the model, we fix the length of arguments to be 80 by truncating the longer arguments and zero-padding the shorter arguments. The word embedding is initialized with pre-trained word vectors using word2vec[1] and unknown words are randomly initialized, and dimensionality setting as 300. After word embedding, we apply dropout and set dropout rate as 0.5. We adopt Momentum [12] with learning rate 0.001 and batch size 90 to train the model.

In the larger-grain linguistic units layer, the convolutional operation uses three groups of 50 filters with filter window sizes of (2,4,8). In the granularities selection layer, the hidden state number of Bi-LSTM is set as 300.

4.3 Overall Performance

We compare our performance with the following state-of-the-art methods, and divide them into two classes. 1) Argument Semantic Learning: **Ji2015** [5] used RNN to encode argument representation and entity that was based on syntax analysis. **Qin2016** [13] took advantage of CNN to extract features of argument pairs. **Qin2017** [14] designed adversarial connective-exploiting networks, which were learned connective features to implicit discourse relation network by adversary between implicit discourse relation network and discriminator. From linguistic point of view, **Lei2018** [7] combined linguistic features. 2) Argument Interactive Learning: **Chen2016** [2] utilized GRN to capture semantic interaction between arguments, and utilized the pooling layer to aggregate interactive information. **Liu2016** [9] designed multiple attention model for adjusting the most relevant information that should be focused on. **Guo2018** [4] proposed interactive attention mechanism to integrate the information of argument pairs into Bi-LSTM so as to get argument representation.

Table 1 shows the overall performance on F1-scores. Lei2018 conducted a comprehensive analysis on PDTB through learning corpus, that their method combined linguistic features such as topic continuity, semantic interaction and attribution. The performance of Lei2018 surpasses our method in Contingency. Our method surpasses the performance of theirs in other relations. All in all, our method that integrating all components obtains state-of-the-art results in terms of Contingency, Expansion and Temporal among the compared models.

The following reasons may explain the performance of our method: 1) In this task, larger-grain linguistic units are semantically cohesive, which contains more useful information than some words. 2) Multi-grain representation consists of word-level representation and larger-grain representation. It can supplement

[1] http://www.code.google.com/p/word2vec

Table 1. The performances of different approaches on the top classes in PDTB in terms of F1-scores(%).

Model	Comp.	Cont.	Expa.	Temp.
Ji2015	35.93	52.78	-	27.63
Qin2016	41.55	57.32	71.50	35.43
Chen2016	40.17	54.76	-	31.32
Liu2016	39.86	54.48	70.43	38.84
Qin2017	40.87	54.56	72.38	36.20
Lei2018	43.24	**57.82**	72.88	29.10
Guo2018	40.35	56.81	72.11	38.65
ours	**45.10**	54.72	**73.3**	**40.18**

richer information than single word. 3) Bidirectional interactive attention mechanism can assign more attention to keywords in argument pairs.

4.4 Our Results and Analysis

In order to verify the effectiveness of the method for larger-grain linguistic units, multi-grain representation learning and the bidirectional interactive attention mechanism, we design five sets of experiments, and the results are shown below.

Table 2. The effects of different components in terms of F1-scores(%).

Model	Comp.	Cont.	Expa.	Temp.
Word-level(basic model)	34.67	42.65	68.73	28.03
LGLU-level	35.56	45.02	69.98	30.47
Word+LGLU	37.70	50.17	70.99	34.99
Word+Interaction	40.16	53.21	70.02	38.07
Word+LGLU+Interaction	45.10	54.72	73.30	40.18

- **Basic Model**: we choose Bi-LSTM as the baseline model. It directly encodes Arg1 and Arg2, and then concatenates them for relation classification. This is an experiment based purely on word-level representation.

- **Larger-Grain Linguistic Units** (LGLU): mainly includes multiple convolutional operations and Bi-LSTM based attention. This is an experiment based on larger-grain representation that directly classify by larger-grain linguistic units embedding.

- **Word+LGLU**: concatenates word information and larger-grain representation into multi-grain representation. So as enhancing argument representation. By means of the comparison between the first experiment and the

third, the performance of the third experiment has been improved among four relations, the F1-scores have increased by 3.03%, 7.52%, 2.26% and 6.96% respectively. For example, example 4) is classified correctly in the third experiment, while wrongly in basic experiment. We infer that our method aggregates "*While not specifically mentioned*" into a whole by filter size 4. It contains more useful information than single word "*mentioned*". If the model classifies via words, "*not*" and "*mentioned*" may be disturbed by "*specifically*". Thus, larger-grain linguistic units are semantically cohesive, and the proposed situation in section 1 is relieved to some extent by our method.

- **Word+Interaction**: when the model learns which compositions are more important in Arg1, the information of Arg2 is used to help model adjustment by calculating interactive attention score (similar operation on Arg2). From the performance, the F1-scores of the fourth experiment are 5.49%, 10.56%, 1.29% and 10.04% higher than the performance of baseline model among four relations. In the first experiment, each word is considered to have the same contribution to infer correct discourse relations, and significant information is not highlighted. Bidirectional interactive attention mechanism relieves the problem. The attention mechanism adjusts the key information of its own through learning information of the other argument, so the classification performance has been substantial improved.

- **Word+LGLU+Interaction**: compared with the third experiment, the fifth one adds bidirectional interactive attention mechanism. Compared with the fourth experiment, the fifth one adds larger-grain linguistic units. Considering example 5), which is classified wrongly in the basic experiment. In the fifth experiment, the larger-grain linguistic unit "*question the authenticity*" and the word "*instead*" are composed as a pair to help correctly classify. The performance exceeds fore four sets experiments. It is further proofed that the effectiveness both of larger-grain linguistic units and bidirectional interactive attention mechanism.

4) [**While not specifically mentioned** *in the FBI charges, dual trading became a focus of attempts to tighten industry regulations*]$_{Arg1}$ [*Implicit=as*] [*Critics contend that traders were putting buying or selling for their own accounts ahead of other traders' customer orders.*]$_{Arg2}$
Relation Type: *Contingency.Cause.Reason*

5) [*scholars* **question the authenticity** *of the Rubens*]$_{Arg1}$ [*Implicit=as*] [*It may have been painted* **instead** *by a Rubens associate.*]$_{Arg2}$
Relation Type: *Contingency.Cause.Reason*

According to the above five experiments, it can be proofed from the improved F1-scores: 1) The larger-grain linguistic units are semantically cohesive, some larger-grain linguistic units contribute more to discourse relation recognition than words. 2) The method of multi-grain representation learning is effective. 3) Each word in an argument has different contribution to judge discourse relation

type. The bidirectional interactive attention mechanism proposed above can help model focus on the information which is effective for the classification.

5 Conclusion

In this paper, we propose a multi-grain representation learning method for implicit discourse relation recognition. The method can automatically obtain larger-grain linguistic units without extracting phrases or chunks by data preprocessing. The multi-grain representation is able to capture complete information of the argument. The bidirectional interactive attention, which is a variant of BiDAF, performs better for information interaction. Experimental results show that the proposed method improves performance among four relation types and obtains comparability compared with the state-of-the-art methods.

Acknowledgments. This research work is supported by National Natural Science Foundation of China (Grants No.61672367, No.61672368, No.61751206). The authors would like to thank the anonymous reviewers for their insightful comments and suggestions. Yu Hong, Professor Associate in Soochow University, is the corresponding author of the paper, whose email address is tianxianer@gmail.com.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6077–6086 (2018)
2. Chen, J., Zhang, Q., Liu, P., Qiu, X., Huang, X.: Implicit discourse relation detection via a deep architecture with gated relevance network. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1726–1735 (2016)
3. Do, Q.X., Chan, Y.S., Roth, D.: Minimally supervised event causality identification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 294–303. Association for Computational Linguistics (2011)
4. 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)
5. Ji, Y., Zhang, G., Eisenstein, J.: Closing the gap: Domain adaptation from explicit to implicit discourse relations. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 2219–2224 (2015)
6. Lei, W., Wang, X., Liu, M., Ilievski, I., He, X., Kan, M.Y.: Swim: A simple word interaction model for implicit discourse relation recognition. In: IJCAI. pp. 4026–4032 (2017)

7. Lei, W., Xiang, Y., Wang, Y., Zhong, Q., Liu, M., Kan, M.Y.: Linguistic properties matter for implicit discourse relation recognition: Combining semantic interaction, topic continuity and attribution. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
8. Litkowski, K.C.: Question-answering using semantic relation triples. In: TREC. Citeseer (1999)
9. Liu, Y., Li, S.: Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. arXiv preprint arXiv:1609.06380 (2016)
10. Pitler, E., Louis, A., Nenkova, A.: Automatic sense prediction for implicit discourse relations in text. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. pp. 683–691. Association for Computational Linguistics (2009)
11. Pitler, E., Nenkova, A.: Revisiting readability: A unified framework for predicting text quality. In: Proceedings of the conference on empirical methods in natural language processing. pp. 186–195. Association for Computational Linguistics (2008)
12. Qian, N.: On the momentum term in gradient descent learning algorithms. Neural networks **12**(1), 145–151 (1999)
13. Qin, L., Zhang, Z., Zhao, H.: A stacking gated neural architecture for implicit discourse relation classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 2263–2270 (2016)
14. Qin, L., Zhang, Z., Zhao, H., Hu, Z., Xing, E.P.: Adversarial connective-exploiting networks for implicit discourse relation classification. arXiv preprint arXiv:1704.00217 (2017)
15. Rutherford, A., Xue, N.: Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 645–654 (2014)
16. Rutherford, A., Xue, N.: Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 799–808 (2015)
17. Tuason, R., Grazian, D., Kondo, G.: Bidaf model for question answering. Table III EVALUATION ON MRC MODELS (TEST SET). Search Zhidao All
18. Xiong, D., Ding, Y., Zhang, M., Tan, C.L.: Lexical chain based cohesion models for document-level statistical machine translation. In: Proceedings of the 2013 conference on empirical methods in Natural Language Processing. pp. 1563–1573 (2013)
19. Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W.W., Salakhutdinov, R., Manning, C.D.: Hotpotqa: A dataset for diverse, explainable multi-hop question answering. arXiv preprint arXiv:1809.09600 (2018)
20. Zhou, L., Li, B., Gao, W., Wei, Z., Wong, K.F.: Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 162–171. Association for Computational Linguistics (2011)
21. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.: Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 207–212 (2016)