A Systematic Investigation of Neural Models for Chinese Implicit Discourse Relationship Recognition

Dejian Li, Man Lan, Yuanbin Wu
School of Computer Science and Technology
East China Normal University, Shanghai, P.R.China
51194506071@stu.ecnu.edu.cn; mlan,ybwu@cs.ecnu.edu.cn

***Abstract*—The Chinese implicit discourse relationship recognition is more challenging than English due to the lack of discourse connectives and high frequency in the text. So far, there is no systematical investigation into the neural components for Chinese implicit discourse relationship. To fill this gap, in this work we present a component-based neural framework to systematically study the Chinese implicit discourse relationship. Experimental results showed that our proposed neural Chinese implicit discourse parser achieves the SOTA performance in CoNLL-2016 corpus.**

***Keywords*-deep learning; Chinese implicit discourse relation recognition; word embedding;**

I. Introduction

Implicit discourse relationship recognition aims to detect the semantic logic relationship (e.g., *Contrast*, *Conjunction*) between consecutive textual units (e.g., clauses, sentences or paragraphs), which is the main challenge of discourse parsing and benefits many downstream NLP tasks such as Sentiment Analysis [1], Machine Translation [2] and Summarization [3], etc. This task is quite challenging due to two reasons: there is no explicit discourse connective (e.g., *because, however*) between textual units (i.e., arguments denoted as `Arg1` and `Arg2`) and implicit discourse relation often occurs in text. For example, almost 40% of the sentences in Penn Discourse Treebank (PDTB)[4] held implicit discourse relations and over 65% in Chinese Discourse Treebank (CDTB)[5].

From the linguistics perspective aspect, the annotation of Chinese discourse relationship differs quite a lot from that in English. Firstly, PDTB has a 3-level hierarchy of multiple relation senses but CDTB has 10 relation senses without hierarchy (as listed in Table I). Secondly, in PDTB `Arg2` is always the text to which the connective is syntactically bound, but in CDTB the text order is dependent on the relation sense rather than the discourse connective. As specified in Table I, in *Causation* relation, the argument of effect is always annotated as `Arg2` no matter the discourse connective is "因为 (because)" or "所以 (thus)". Thirdly, in PDTB the discourse relations are annotated within one paragraph while in CDTB the implicit relation can be hold across paragraphs. Fourthly, since the sentences in Chinese are often short (may share the same subject), leading to a large proportion of *Conjunction* relations in Chinese.

Table I
The order of arguments for each sense of discourse relations in CDTB.

Sense	Arg1	Arg2
Alternative		或者/or
Causation	因为/because	所以/so
Conditional	如果/if	就/then
Conjunction		而且/and
Contrast	虽然/although	但是/but
Expansion	综上所述/in conclusion	例如/for example
Progression	不仅/not only	还/but also
Progression	通过/through	还/-
Restatement		换言之/in other words
Temporal	在... 之后/text order	在... 之前/text order

Recent studies on Chinese implicit discourse relationship adopted deep learning methods. [6] first examined several unsupervised word representations (e.g., one-hot word pair, Brown word clusters and simple word embedding) and confirmed the effectiveness of word embeddings. Later, [7] and [8] explored neural models by adopting word2vec embedding and element-wise *pooling* functions (i.e., *max*, *sum* and *mean*) for sentence representation but they neglected the relationship interaction between two arguments. Furthermore, [8] and [9] used an self-attention BiLSTM to derive sentence representation and demonstrated that modeling two arguments as a joint sequence outperforms previous word order-agnostic approaches.

However, the studies described above leave two open questions. First, several recent word embeddings (e.g., GloVe, ELMo, BERT) have been reported supreme performance in many NLP tasks. We would like to examine their performance for Chinese implicit discourse recognition. Second, the discourse relationship between two arguments is supposed to be more complicated than simple concatenation or self-attention operations on two arguments. We state that the discourse relation between arguments is represented by the interaction between two arguments rather than only simple or separate operations of two arguments.

To address these questions, we present a systematic investigation work to deeply analyze the influence of neural models. Our main contributions are summarized as follows.

- To our knowledge, this is the first framework to systematically investigate neural models for Chinese implicit discourse relationship recogni-

tion. We present a component-based deep learning architecture, which consists of four independent components and each of them have multiple implementations.
- Extensive experiments in benchmark CoNLL-2016 corpus are conducted to demonstrate the efficacy and effectiveness of our model. Our proposed neural model outperforms the state-of-the-art models by average 1% in accuracy.

II. Related Work

Discourse relationship recognition has attracted a lot of research interests in these years. The recognition of explicit discourse relationship reaches 93% accuracy only by using discourse connectives [10], but the performance of implicit discourse relationship recognition is always poor due to the lack of discourse connectives, which is the bottleneck of the whole discourse parser. Earlier Researchers adopted traditional NLP methods to design and extract complex features with expert knowledge. [11] adopted an aggregated approach to word pairs and [12] employed Brown word clusters.These methods perform badly in generalization.

With the development of deep learning in NLP, researchers began to use the deep learning method to recognize implicit discourse relationship. For example, [6] first compared different unsupervised word representations including standard one-hot word pair representations, low-dimensional representations based on Brown clusters and word embedding. They demonstrated the effectiveness of the word embedding. The studies using deep learning methods are divided into two lines in general. One research line is to learn from explicit discourse relationship or other languages. [13], [14], [15] tried to expand the implicit training dataset with the help of discourse connectives. In order to make full use of the connectives in explicit data, [16] used connective-based word representations and [17] learned discourse-specific word embedding from massive explicit data. [18] presented their implicit network to learn from another neural network which has access to connectives. Unlike those above, [19] used bilingually-constrained synthetic implicit data for implicit discourse relation recognition. The other line focuses on the expression of words and the structure of the model. For example, [20], [21] used word2vec word embedding and Convolutional Neural Network (CNN) to determine the senses. [22] used CNN to model argument pairs with GloVe word embedding and multi-task learning system. [23] combined the word2vec and their proposed event embedding. [24] combined the context information into word embedding which is context-aware character-enhanced embeddings. Regarding to model structure, [25], [26] adopted gated network to calculate the relevance score between two arguments. [27] employed new network structure TreeLSTM to model the sentences.

However, all above studies focused on English corpus and there is not much studies on Chinese corpus. [7], [8] explored feedforward and LSTM for this task. [9] used BiLSTM to model the sentences. To alleviate the shortage of labeled data, [19] designed a multi-task neural network model to use their bilingually-constrained synthetic implicit data as additional data.

III. Chinese Implicit Discourse Relationship Parser

We present a component-based neural framework for Chinese implicit discourse relationship recognition, consisting of four independent components. Figure 1 depicts the architecture of Chinese implicit discourse relationship parser.

A. Word Embedding Layer

Word embedding is the first and crucial step in deep learning framework, which transforms the natural language into word vector as the input of the neural network. To do so, we convert each word $\boldsymbol{w}$ into a word vector $\boldsymbol{x} \in \mathbb{R}^{d_w}$, where the d_w is the dimension of the word vector. Let $\boldsymbol{x}_i^1$ $\left(\boldsymbol{x}_i^2\right)$ be the i-th word vector in Arg-1(Arg-2), then the two discourse arguments are represented as:

$$Arg\text{-}1 : \left[\boldsymbol{x}_1^1, \boldsymbol{x}_2^1, \cdots, \boldsymbol{x}_{L_1}^1\right] \tag{1}$$
$$Arg\text{-}2 : \left[\boldsymbol{x}_1^2, \boldsymbol{x}_2^2, \cdots, \boldsymbol{x}_{L_2}^2\right] \tag{2}$$

where Arg-1(Arg-2) has $L_1(L_2)$words.

Generally, the word embeddings are pre-trained on large corpus and supposed to contain latent semantic and syntactic information. In recent years several supreme word embeddings have been presented by researchers. To examine their different effectiveness in word conversion, we choose two types of pre-trained word vector models, i.e., context-free models and contextual models.

Context-free models generate a word embedding representation for each word in the vocabulary, without regard to the context of this word in specific arguments. Word2vec [28] and GloVe [29] are two widely used context-free models. Word2vec uses local text controlled by small window size from large corpus to train the word vector. While GloVe trains on aggregated global word co-occurrence statistics from the corpus.

Contextual models generate word representation for each word based on its context words in the sentence. Usually contextual models aims to obtain language model rather than word embedding. For specific sentence, it gets contextual word representation base on language model. Here we choose ELMo and BERT models as follows.

ELMo [30] is a deep contextualized word representation, which is learned as the internal states of a deep bidirectional language model (biLM).

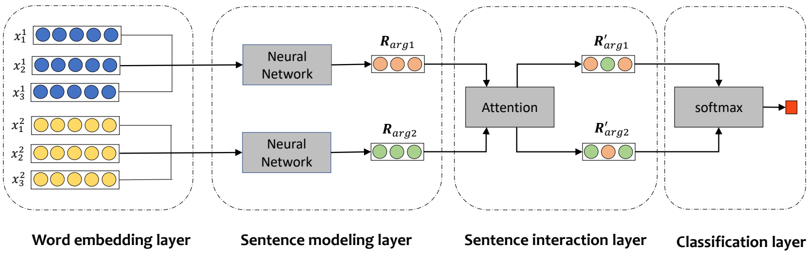

Figure 1. Architecture of our implicit discourse relationship parser system.

BERT[1] means Bidirectional Encoder Representations from Transformers, which is learned from unlabeled text by jointly conditioning on both left and right context in all layers [31].

The two above contextual models are different in some aspects. Firstly, ELMo uses LSTM to encode word while BERT uses transformers. Besides, they have different bi-direction implementations. In ELMo, the forward and the backward direction are simply aggregated and the bidirectional is separate from training. But the bidirectional of the BERT is integrated into the training process.

B. Sentence Modeling Layer

Each argument is transformed into a word vector matrix as shown in Formula (1) and (2) from the word embedding layer. To achieve the semantic representation for each argument, we adopt three sentence modeling methods, i.e., Long Short Term Memory (LSTM)[32], Bi-directional Long Short Term Memory (BiLSTM)[33] and Convolutional Neural Network (CNN).

Given the two argument representations as shown in Formula (1) and (2), the LSTM computes the state sequence $[\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_L]$ for each time step i using the following formulas:

$$\boldsymbol{i}_i = \sigma(\boldsymbol{W}_i[\boldsymbol{x}_i, \boldsymbol{h}_{i-1}] + \boldsymbol{b}_i) \tag{3}$$
$$\boldsymbol{f}_i = \sigma(\boldsymbol{W}_f[\boldsymbol{x}_i, \boldsymbol{h}_{i-1}] + \boldsymbol{b}_f) \tag{4}$$
$$\boldsymbol{o}_i = \sigma(\boldsymbol{W}_o[\boldsymbol{x}_i, \boldsymbol{h}_{i-1}] + \boldsymbol{b}_o) \tag{5}$$
$$\tilde{\boldsymbol{c}}_i = tanh(\boldsymbol{W}_c[\boldsymbol{x}_i, \boldsymbol{h}_{i-1}] + \boldsymbol{b}_c) \tag{6}$$
$$\boldsymbol{c}_i = \boldsymbol{i}_i \odot \tilde{\boldsymbol{c}}_i + \boldsymbol{f}_i \odot \boldsymbol{c}_{i-1} \tag{7}$$
$$\boldsymbol{h}_i = \boldsymbol{o}_i \odot tanh(\boldsymbol{c}_i) \tag{8}$$

where σ denotes the *sigmoid* function and $\odot$ denotes element-wise multiplication.

Unlike LSTM using information only from past, BiLSTM gets the information from both past and future directions. At each position i of the sequence, we obtain two states $\overrightarrow{\boldsymbol{h}}_i$ and $\overleftarrow{\boldsymbol{h}}_i$, where $\overrightarrow{\boldsymbol{h}}_i, \overleftarrow{\boldsymbol{h}}_i \in \mathbb{R}^{d_h}$. Then we concatenate them to get the intermediate state, i.e. $\boldsymbol{h}_i = [\overrightarrow{\boldsymbol{h}}_i, \overleftarrow{\boldsymbol{h}}_i]$. After that, we sum up the states in sequence $[\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_L]$ to get the representations of *Arg*-1 and *Arg*-2 as follows:

$$\boldsymbol{R}_{Arg_1} = \sum_{i=1}^{L_1} \boldsymbol{h}_i^1 \tag{9}$$
$$\boldsymbol{R}_{Arg_2} = \sum_{i=1}^{L_2} \boldsymbol{h}_i^2 \tag{10}$$

In CNN model, we use $\boldsymbol{Arg}[i:j]$ to represent the sub-matrix of $\boldsymbol{Arg}$ from row i to row j. A convolution involves a filter $\mathbf{w} \in \mathbb{R}^{h \times d}$ (h is the height of filter and d is the dimensionality of the word vector) The output sequence o_i of the convolution operator is obtained by repeatedly applying the filter on sub-matrices of $\boldsymbol{Arg}$ as follows:

$$o_i = \mathbf{w} \cdot \mathbf{Arg}[i : i + h - 1] \tag{11}$$

where i=1...$s - h + 1$. A bias term $b \in \mathbb{R}$ and an activation function f are added to each o_i to compute the feature map c_i for this filter:

$$c_i = f\,(o_i + b) \tag{12}$$

Then we use *max pooling* operation to get the representation of the argument:

$$\boldsymbol{R_{arg}} = max\{c_i\} \tag{13}$$

C. Sentence Interaction Layer

Unlike sentence modeling to get semantic representation for each sentence, the sentence interaction aims to learn the relationship representation between two arguments rather than one single argument. We choose four ways to make arguments concatenate with each other, resulting in an interrelated representation of the two arguments. These methods based on the attention and the self-attention mechanism as follows.

- Attention: Perform attention operations on two argument vectors respectively and then concatenate them together
- Con-self-attention: Concatenate two argument vectors first and then perform self-attention operation on the concatenated vector
- Self-attention-con: Perform self-attention operation on two argument vectors respectively and then concatenate them together
- Attention-mlp: Perform attention interact operations on two sentence vectors respectively and

[1]https://github.com/google-research/BERT\#pre-trained-models

then feed their concatenation into Multi-Layer Perceptron (MLP)

Through the above four interactions, the separate $\boldsymbol{R_{Arg1}}$, $\boldsymbol{R_{Arg2}}$ become joint pair representation $\boldsymbol{R_{pair}}$ which contains the overall information of the two arguments.

D. Classification Layer

Finally we feed the result of joint representation of the arguments $\boldsymbol{R}_{pair}$ into a full-connected *softmax* layer to predict the implicit discourse sense.

IV. Experiment

A. Dataset

We perform our experiments on the CDTB corpus[2]. To make comparison with previous work, we use the data provided by the CoNLL-2016, which is adapted from the CDTB corpus and has been a benchmark corpus for study. Following previous work, we use the accuracy to evaluate the performance of models. Table II shows the distributions of Chinese and English corpus in CoNLL-2016.

Table II
The distributions of discourse relationship types in CoNLL-2016 English and Chinese corpus.

	English		Chinese	
	amount	percent(%)	amount	percent(%)
Explicit	18,459	45.5	2,398	21.75
Implicit	16,053	39.5	7,238	65.66
EntRel	5,210	12.8	1,219	11.06
AltLex	624	1.5	223	2.02
NoRel	254	0.6	0	0
Total	40,600	100	11,023	100

Clearly, there is more implicit data in Chinese than in English. We follow the previous work in [20] and combine the non-Explicit (i.e.,Implicit, EntRel and AltLex) dataset as implicit samples, which makes the implicit discourse relationship recognition more challenging. The Chinese discourse relationship is divided into 9 categories. Table III lists the sense distribution breakdown of Chinese non-Explicit discourse.

Table III
Conll-2016 Chinese non-Explicit sense distribution.

Sense Label	Training	Development	Test
Conjunction	5,196	189	228
Expansion	1,228	49	40
EntRel	1,098	50	71
Causation	260	12	11
Purpose	79	2	6
Contrast	72	3	1
Temporal	36	0	1
Conditional	32	1	1
Progression	14	0	0

[2]https://catalog.ldc.upenn.edu/LDC2013T21

Table V
Comparisons of Accuracy(%) for sentence interactions.

	word2vec +BiLSTM	$ELMo_2$ + BiLSTM	$BERT_{single}$ + CNN
without Attention	70.19	72.70	**74.09**
Attention	68.80	68.52	70.75
Con-self-attention	70.75	65.74	71.30
Self-attention-con	72.98	70.20	70.75
Attention-mlp	70.47	64.90	69.63

B. Experiment Setup

We employ Adam optimization [34] using the cross-entropy loss function. In CNN model, [20] choose filter window size $(1, 3, 5)$ to represent the *unigram, trigram* and *5-gram* features in sentence. We following their choice because we test all the sub-set of $(1, 3, 5, 7)$ and found $(1, 3, 5)$ achieves the optimal performance. Following [35], we set hidden size as 50 in LSTM and BiLSTM. We set epochs as 50, batch size as 64, learning rate as 0.001 and dropout as 0.5.

In the word embedding layer, learning from the positive correlation between the vector dimension and expression ability in English, we train the 300 dimensions of word2vec and GloVe vector on Tagged Chinese Gigaword[3]. In the contextual embedding model, we get the three-layers ELMo representations by the tools provided by allennlp[4] to train the contextual representation of words. As for BERT, we use the pre-trained Chinese model offered by Google, which is 12-layer, 768-hidden, 12-heads, 110M parameters. Both single sentence model and sentence pairs model are used in our experiment for BERT.

C. Results and Discussion

We evaluate our component-based model from different aspects, i.e., the contextual and context-free word embeddings, three sentence modeling methods and two relationship interaction strategies. Table IV and Table V reported the experimental results on test dataset. Next we analyze the performance of different components.

Table IV
Comparisons of Accuracy(%) for different word embeddings and sentence modeling methods.

	word2vec	*GloVe*	*ELMo* 1	2	3	*BERT* single	pairs
LSTM	67.68	70.47	70.31	71.59	67.41	67.69	66.57
BiLSTM	70.19	71.30	71.03	72.70	70.47	66.30	68.24
CNN	70.75	70.20	69.92	71.59	72.42	**74.09**	70.47

1) **word embeddings:** To examine the impact of different word embeddings and sentence modeling, we did not involve the operation of sentence interaction. From Table IV, we find BERT_{single} with CNN achieves the best performance (74.09% in accuracy). However,

[3]https://catalog.ldc.upenn.edu/LDC2007T03
[4]https://github.com/allenai/allennlp

comparing other embeddings, BERT performs not stable enough considering the worst is BERT_{single} with BiLSTM (66.30% in accuracy). The possible reason may be that BERT generates the single Chinese character embeddings rather than the word embeddings and thus the simple addition of two characters cannot be equal to the real word embedding. Furthermore, the ELMo_{second} performs well and stable comparing other embeddings.

2) **sentence modeling:** The performance of different sentence modeling methods cannot be summarized in one sentence. When in combination with different word embeddings, the performance of sentence modeling methods changes a lot. For example, comparing CNN with BiLSTM and LSTM, we find the bidirectional information is helpful in discourse relationship recognition when in combination with most word embeddings. However, CNN with BERT_{single} achieves the best performance among all setting. This indicates that not only word embedding but with their combination makes contribution to performance.

3) **sentence interactions:** Furthermore, we examine the performance of different sentence interactions along with the selected combination of word embedding and sentence modeling as shown in Table IV. From Table V we see that the sentence interactions did not perform well as we expected. Given word embedding and sentence modeling, the sentence interaction without attention outperforms other attention strategies. This is surprising as we think the diverse discourse relations is complex and supposed to be represented by complex operations rather than simple concatenation. To give a deep analysis of this phenomenon, we dive in to the Chinese implicit discourse corpus. We find that among all senses, *Conjunction* is a very common category (63.5%) as it is the default category when the relationship is hard to judge[36]. Since the two arguments in *Conjunction* are often flexible and varied in structure and content, the interaction relationship between the two arguments may not be effectively captured by the attention mechanisms proposed in this work. This opens a future study for interaction relationship representation.

Finally, Table VI shows the comparison of our best model with the recent state-of-the-art systems on CoNLL-2016 for multi-class classification. All these systems use gold standard argument pairs. Again, our model BERT_{single} with CNN achieves the best performance (74.09% in accuracy) and outperforms the state-of-the-art performance. This indicates the effectiveness of our proposed model.

Table VI
Comparisons of our best model with recent systems on CoNLL-2016 Chinese non-Explicit dataset, accuracy(%).

	Development Set	Test Set
Wang and Lan (2016) [20]	**73.53**	72.42
Rutherford and Xue (2016) [37]	71.57	67.41
Schenk et al. (2016)[7]	70.59	71.87
Rönnqvist et al. (2017)[9]	-	73.01
Ours($BERT_{single}$+CNN)	72.54	**74.09**

V. Conclusion

In this paper, we present a component-based neural framework to investigate the neural components for Chinese implicit discourse relationship recognition. Different word embeddings, sentence modeling methods and relationship interaction strategies are extensively explored. The experimental results showed that it is not one component does matter but their combination makes contribution to performance improvement. Besides, our proposed model achieves the SOTA performance in CoNLL-2016 Chinese corpus.

References

[1] B. Yang and C. Cardie, "Context-aware learning for sentence-level sentiment analysis with posterior regularization," in *Proceedings of the 52nd ACL*, 2014, pp. 325–335.

[2] J. J. Li, M. Carpuat, and A. Nenkova, "Assessing the discourse factors that influence the quality of machine translation," in *Proceedings of the 52nd ACL*, 2014, pp. 283–288.

[3] X. Wang, Y. Yoshida, T. Hirao, K. Sudoh, and M. Nagata, "Summarization based on task-oriented discourse parsing," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 1358–1367, 2015.

[4] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. K. Joshi, and B. L. Webber, "The penn discourse treebank 2.0." in *LREC*, 2008.

[5] Y. Zhou and N. Xue, "Pdtb-style discourse annotation of chinese text," in *Proceedings of the 50th ACL: Long Papers-Volume 1*, 2012, pp. 69–77.

[6] C. Braud and P. Denis, "Comparing word representations for implicit discourse relation classification," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2201–2211.

[7] N. Schenk, C. Chiarcos, K. Donandt, S. Rönnqvist, E. Stepanov, and G. Riccardi, "Do we really need all those rich linguistic features? a neural network-based approach to implicit sense labeling," in *Proceedings of the CoNLL-16 shared task*, 2016, pp. 41–49.

[8] A. Rutherford, V. Demberg, and N. Xue, "A systematic study of neural discourse models for implicit discourse relation," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017, pp. 281–291.

[9] S. Rönnqvist, N. Schenk, and C. Chiarcos, "A recurrent neural model with attention for the recognition of Chinese implicit discourse relations," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017, pp. 256–262.

[10] E. Pitler and A. Nenkova, "Using syntax to disambiguate explicit discourse connectives in text," in *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, 2009, pp. 13–16.

[11] O. Biran and K. McKeown, "Aggregated word pair features for implicit discourse relation disambiguation," 2013.

[12] A. Rutherford and N. Xue, "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*, 2014, pp. 645–654.

[13] C. Braud and P. Denis, "Combining natural and artificial examples to improve implicit discourse relation identification," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 1694–1705.

[14] A. Rutherford and N. Xue, "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*, 2015, pp. 799–808.

[15] Y. Ji, G. Zhang, and J. Eisenstein, "Closing the gap: Domain adaptation from explicit to implicit discourse relations," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2219–2224.

[16] C. Braud and P. Denis, "Learning connective-based word representations for implicit discourse relation identification," 2016.

[17] C. Wu, X. Shi, Y. Chen, J. Su, and B. Wang, "Improving implicit discourse relation recognition with discourse-specific word embeddings," in *Proceedings of the 55th ACL, year=2017*.

[18] L. Qin, Z. Zhang, H. Zhao, Z. Hu, and E. Xing, "Adversarial connective-exploiting networks for implicit discourse relation classification," in *Proceedings of the 55th Annual Meeting of ACL*, 2017.

[19] C. Wu, Y. Chen, Y. Huang *et al.*, "Bilingually-constrained synthetic data for implicit discourse relation recognition," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2306–2312.

[20] J. Wang and M. Lan, "Two end-to-end shallow discourse parsers for english and chinese in conll-2016 shared task," in *Proceedings of the CoNLL-16 shared task*, 2016.

[21] L. Qin, Z. Zhang, and H. Zhao, "Shallow discourse parsing using convolutional neural network," *Proceedings of the CoNLL-16 shared task*, pp. 70–77, 2016.

[22] Y. Liu, S. Li, X. Zhang, and Z. Sui, "Implicit discourse relation classification via multi-task neural networks," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[23] M. L. Pacheco, I.-T. Lee, X. Zhang, A. K. Zehady, P. Daga, D. Jin, A. Parolia, and D. Goldwasser, "Adapting event embedding for implicit discourse relation recognition," *Proceedings of the CoNLL-16 shared task*, pp. 136–142, 2016.

[24] L. Qin, Z. Zhang, and H. Zhao, "Implicit discourse relation recognition with context-aware character-enhanced embeddings," in *Proceedings of COLING 2016: Technical Papers*, pp. 1914–1924.

[25] ——, "A stacking gated neural architecture for implicit discourse relation classification," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2263–2270.

[26] J. Chen, Q. Zhang, P. Liu, X. Qiu, and X. Huang, "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*, 2016, pp. 1726–1735.

[27] Y. Wang, S. Li, J. Yang, X. Sun, and H. Wang, "Tag-enhanced tree-structured neural networks for implicit discourse relation classification," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, 2017, pp. 496–505.

[28] T. Mikolov, I. Sutskever, K. Chen, Corrado, and G. S., "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013.

[29] Pennington, Jeffrey, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on EMNLP*, 2014.

[30] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proc. of NAACL*, 2018.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, pp. 1735–1780.

[33] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, pp. 602–610.

[34] K. D. P and B. J. Adam, "Adam: A method for stochastic optimization," 2014.

[35] M. Lan, J. Wang, Y. Wu, Z.-Y. Niu, and H. Wang, "Multi-task attention-based neural networks for implicit discourse relationship representation and identification," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1299–1308.

[36] e. a. Xue, Nianwen, "Conll 2016 shared task on multilingual shallow discourse parsing." in *Proceedings of the CoNLL-16 shared task (2016)*, 2016, pp. 1–19.

[37] A. Rutherford and N. Xue, "Robust non-explicit neural discourse parser in english and chinese," *Proceedings of the CoNLL-16 shared task*, pp. 55–59, 2016.