Implicit Discourse Relation Recognition using Neural Tensor Network with Interactive Attention and Sparse Learning

Fengyu Guo[1,2,*], **Ruifang He**[1,2,*,†], **Di Jin**[1,2], **Jianwu Dang**[1,2,3], **Longbiao Wang**[1,2], **Xiangang Li**[4]
[1]School of Computer Science and Technology, Tianjin University, Tianjin, China.
[2]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin, China.
[3]Japan Advanced Institute of Science and Technology, Ishikawa, Japan.
[4]AI Labs, Didi Chuxing, Beijing, China.
{fengyuguo,rfhe,jindi,longbiao_wang}@tju.edu.cn
jdang@jaist.ac.jp, lixiangang@didichuxing.com

Abstract

Implicit discourse relation recognition aims to understand and annotate the latent relations between two discourse arguments, such as temporal, comparison, etc. Most previous methods encode two discourse arguments separately, the ones considering pair specific clues ignore the bidirectional interactions between two arguments and the sparsity of pair patterns. In this paper, we propose a novel Neural **T**ensor Network framework with **I**nteractive **A**ttention and **S**parse **L**earning (TIASL) for implicit discourse relation recognition. (1) We mine the most correlated word pairs from two discourse arguments to model pair specific clues, and integrate them as interactive attention into argument representations produced by the bidirectional long short-term memory network. Meanwhile, (2) the neural tensor network with sparse constraint is proposed to explore the deeper and the more important pair patterns so as to fully recognize discourse relations. The experimental results on PDTB show that our proposed TIASL framework is effective.

1 Introduction

Discourse relation describes how two adjacent text units (e.g. clauses, sentences, and larger sentence groups), called arguments, named $Arg1$ and $Arg2$, are connected semantically, such as temporally, causally, etc. Yet implicit discourse relation recognition without explicit connectives (Pitler et al., 2008), which needs to infer the relation from specific context, is still a challenging problem. It can be used in text summarization(Gerani et al., 2014), conversation system (Higashinaka et al., 2014) and so on.

Previous researches mainly include (1) traditional feature-based models and (2) neural network based models. Most feature-based models adopt various linguistic features (such as polarity, word pairs, and position information, etc.) and design complicated rules to recognize implicit discourse relations (Pitler et al., 2009; Zhou et al., 2010; Braud and Denis, 2015). They can not fully use the local and the global context, and the human cost is huge. Neural network based models get the better argument representations and more precisely capture discourse relations (Braud and Denis, 2015; Zhang et al., 2015; Liu et al., 2016). However, they encode two discourse arguments separately, and ignore pair specific clues. The further researches adopt the different hybrid neural models (Chen et al., 2016; Lei et al., 2017) and attention mechanism (Cai and Zhao, 2017) to mine the semantic interactions of argument pairs. Yet, they ignore the bidirectional interactions between two arguments during the representation stage since there is asymmetry from the perspective of human-like reading strategy. And the sparsity of word pair patterns indicating discourse relation is neither considered.

Therefore, a novel neural **T**ensor network model with **I**nteractive **A**ttention and **S**parse **L**earning is proposed for implicit discourse relation recognition, namely TIASL. We imitate the human-like reading strategy, and model the relatedness between two discourse arguments as a kind of interactive attention from bidirectional aspects. It is added into the argument representations with a bidirectional Long Short-Term Memory network (Bi-LSTM), and then plugged in neural tensor network (NTN) with l_1 reg-

* Equal contribution.

† Corresponding author.

Proceedings of the 27th International Conference on Computational Linguistics, pages 547–558
Santa Fe, New Mexico, USA, August 20-26, 2018.

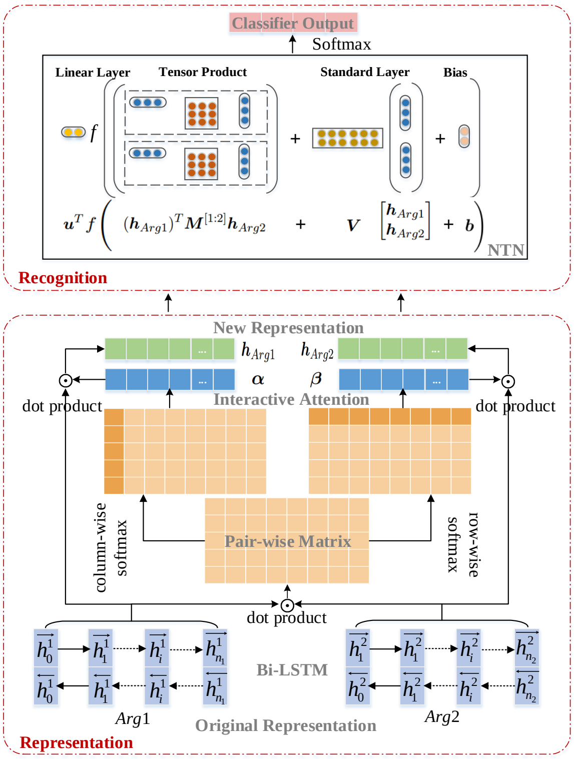


Figure 1: The TIASL framework.

ularization. This helps to mine the different aspects of semantic interactions between two arguments and select the important and the informative word pair patterns.

Our main contributions are as follows:

- Propose a novel TIASL framework from the perspectives of the human-like bidirectional reading strategy and the sparsity of word pair patterns;
- Encode the discourse arguments by the Bi-LSTM with interactive attention for implicit discourse relation recognition;
- Use neural tensor network with sparse constraint to capture the deeper and the more indicative pair patterns;
- Experimental results on PDTB show that our TIASL model is effective.

2 The Proposed Method

We formalize implicit discourse relation recognition as a classification problem. The proposed TIASL framework is shown in Figure 1. The main steps include (1) discourse argument representations with interactive attention based on Bi-LSTM and (2) sparse pair pattern selection and implicit discourse relation recognition.

2.1 Discourse Argument Representations with Interactive Attention

Attention mechanism has achieved great success in image recognition, which is based on the visual attention principle found in humans. Recently, it is widely adopted in many NLP tasks. Inspired by (Herzog et al., 2016), we imitate the human-like bidirectional reading strategy, and propose an interactive attention mechanism to enhance discourse argument representations.

For the original representations of discourse arguments shown in Figure 1, we first associate each word w in the vocabulary with a vector representation $\boldsymbol{x}_w \in \mathbb{R}^d$, where d is the dimension of the embeddings.

Since each argument is viewed as a sequence of word vectors, let $\boldsymbol{x}_i^1(\boldsymbol{x}_i^2)$ be the i-th word vector in *Arg*1 (*Arg*2), thus the arguments in a discourse relation are expressed as,

$$Arg1 : [\boldsymbol{x}_1^1, \boldsymbol{x}_2^1, ..., \boldsymbol{x}_{n1}^1], \qquad Arg2 : [\boldsymbol{x}_1^2, \boldsymbol{x}_2^2, ..., \boldsymbol{x}_{n2}^2].$$

where *Arg*1 (*Arg*2) has n_1 (n_2) words.

2.1.1 The Basic Bi-LSTM

Long Short-Term Memory network (LSTM) (Hochreiter and Schmidhuber, 1997) is a variant of recurrent neural network. Considering that it can model long-term dependencies and encode context information, we use it in the basic argument representation. Given the word representations of two arguments as we just described, the LSTM computes the state sequence for each position t using the following equations:

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i[\boldsymbol{x}_t, \boldsymbol{h}_{t-1}] + \boldsymbol{b}_i), \tag{1}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f[\boldsymbol{x}_t, \boldsymbol{h}_{t-1}] + \boldsymbol{b}_f), \tag{2}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o[\boldsymbol{x}_t, \boldsymbol{h}_{t-1}] + \boldsymbol{b}_o), \tag{3}$$

$$\tilde{\boldsymbol{c}}_t = \tanh(\boldsymbol{W}_c[\boldsymbol{x}_t, \boldsymbol{h}_{t-1}] + \boldsymbol{b}_c), \tag{4}$$

$$\boldsymbol{c}_t = \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t + \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1}, \tag{5}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t). \tag{6}$$

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

Since LSTM only considers the context from the previous, we utilize a bidirectional LSTM (Bi-LSTM) preserving both history and future information. Therefore, at each position t of the sequence, we can obtain two representations $\overrightarrow{\boldsymbol{h}_t}$ and $\overleftarrow{\boldsymbol{h}_t}$. Then we concatenate them to get the intermediate state $\boldsymbol{h}_t = [\overrightarrow{\boldsymbol{h}_t}, \overleftarrow{\boldsymbol{h}_t}]$. For *Arg*1 and *Arg*2, we encode them into the contextual representations by Bi-LSTM. That is, $\boldsymbol{h}_i^1 = [\overrightarrow{\boldsymbol{h}_i^1}, \overleftarrow{\boldsymbol{h}_i^1}]$ and $\boldsymbol{h}_j^2 = [\overrightarrow{\boldsymbol{h}_j^2}, \overleftarrow{\boldsymbol{h}_j^2}]$ are the intermediate states of i-th word in *Arg*1 and j-th word in *Arg*2 respectively, where $\overrightarrow{\boldsymbol{h}_i^1}, \overleftarrow{\boldsymbol{h}_i^1}, \overrightarrow{\boldsymbol{h}_j^2}, \overleftarrow{\boldsymbol{h}_j^2} \in \mathbb{R}^d$.

Separately encoding arguments with Bi-LSTM could not reflect the semantic between two discourse arguments in a discourse relation. In order to fully use their semantic connections, we explore a novel argument representation.

2.1.2 Model the Asymmetry of Reciprocal Attention on Discourse Arguments

Herzog et al. (2016) proposed the two stage model of visual perception which indicated that people‘s image recognition includes two stages: collecting information and understanding information. In daily life, we have a similar feeling intuitively during reading: a more reasonable strategy is that people may read two discourse arguments back and forth, and find some relevant and informative clues helpful to judge the discourse relation. Due to the different reading order of two arguments, people may get the different focused information and thus have the different decisions. Therefore, we model the reciprocal attention on discourse arguments from two directions.

Firstly, we calculate semantic connections between word pairs in two arguments as a pair-wise matrix shown in Eq.(7), which indicates the relevant score of i-th *Arg*1 word and j-th *Arg*2 word by dot product of their hidden representations.

$$\boldsymbol{S}(i, j) = (\boldsymbol{h}_i^1)^T \cdot \boldsymbol{h}_j^2. \tag{7}$$

where $\boldsymbol{S} \in \mathbb{R}^{n_1 \times n_2}$, n_1 and n_2 are the lengths of *Arg*1 and *Arg*2, respectively.

Secondly, as the concerns of the forward and the reverse reading order are asymmetric when judging the relation of two discourse arguments. For each word in *Arg*2, we apply a column-wise softmax function on the pair-wise matrix $\boldsymbol{S}$ to get a probability distribution $\boldsymbol{\alpha}_t$ over *Arg*1, shown in Eq.(8). Similarly, we conduct a row-wise softmax function to get $\boldsymbol{\beta}_t$ over *Arg*2 when considering one *Arg*1 word, shown in Eq.(9). We denote $\boldsymbol{\alpha}_t \in \mathbb{R}^{n_1}$ as *Arg*2-to-*Arg*1 attention, and $\boldsymbol{\beta}_t \in \mathbb{R}^{n_2}$ as *Arg*1-to-*Arg*2 attention at position t, which are named **interactive attention**.

$$\boldsymbol{\alpha}_t = softmax(\boldsymbol{S}(1,t), ..., \boldsymbol{S}(n_1,t)), \tag{8}$$

$$\boldsymbol{\beta}_t = softmax(\boldsymbol{S}(t,1), ..., \boldsymbol{S}(t,n_2)). \tag{9}$$

where $\boldsymbol{\alpha}_t = [\boldsymbol{\alpha}_t^1, \boldsymbol{\alpha}_t^2, ..., \boldsymbol{\alpha}_t^{n_1}]$, $\boldsymbol{\alpha}_t^i$ means the attention value of i-th word in *Arg*1 at position t. Likewise, $\boldsymbol{\beta}_t = [\boldsymbol{\beta}_t^1, \boldsymbol{\beta}_t^2, ..., \boldsymbol{\beta}_t^{n_2}]$, $\boldsymbol{\beta}_t^j$ is the attention value of j-th word in *Arg*2 at position t.

In order to exploit the overall influence information to represent semantic connection of two discourse arguments, we average all the $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ to get the final attention of *Arg*1 and *Arg*2.

$$\boldsymbol{\alpha} = \frac{1}{n_2}\sum_{t=1}^{n_2}\boldsymbol{\alpha}_t, \qquad \boldsymbol{\beta} = \frac{1}{n_1}\sum_{t=1}^{n_1}\boldsymbol{\beta}_t. \tag{10}$$

The new argument representations integrating argument context and interactive attention are shown as Eq.(11), which reflect the human-like bidirectional reading strategy to some extent.

$$\boldsymbol{h}_{Arg1} = \boldsymbol{h}^1\boldsymbol{\alpha}, \qquad \boldsymbol{h}_{Arg2} = \boldsymbol{h}^2\boldsymbol{\beta}. \tag{11}$$

2.2 Sparse Pair Pattern Selection and Discourse Relation Recognition

Observations show that there are some pair patterns in a discourse relation. Once we detect these interactions expressing pair patterns, discriminating discourse relation is obvious. However, how to represent and select this kind of interaction is a problem.

2.2.1 Neural Tensor Network

Conventional methods to measure the relevance between two arguments includes bilinear model (Jenatton et al., 2012), and single layer neural networks (Collobert and Weston, 2008), etc. These methods could hardly model the complex and informative interactions. Success in knowledge graph (Socher et al., 2013a) shows that tensor can model multiple interactions in data. Therefore, we further employ a tensor layer to mine the deeper semantic interactions based on the new argument representations so as to recognize implicit discourse relations.

Tensor is a geometric object that describes relations between vectors, scalars, and others. It can be represented as a multi-dimensional array of numerical values. Following the NTN (Socher et al., 2013a; Pei et al., 2014), we utilize a 3-way tensor $\boldsymbol{M}^{[1:k]} \in \mathbb{R}^{d_h \times d_h \times k}$ to model the interactions shown in Eq.(12).

$$g(\boldsymbol{h}_{Arg1}, \boldsymbol{h}_{Arg2}) = \boldsymbol{u}^T f\left((\boldsymbol{h}_{Arg1})^T \boldsymbol{M}^{[1:k]} \boldsymbol{h}_{Arg2} + \boldsymbol{V}\begin{bmatrix}\boldsymbol{h}_{Arg1}\\ \boldsymbol{h}_{Arg2}\end{bmatrix} + \boldsymbol{b}\right). \tag{12}$$

where f is a standard nonlinearity applied element-wise, $\boldsymbol{M}^{[1:k]} \in \mathbb{R}^{d_h \times d_h \times k}$ is a tensor and the bilinear tensor product $(\boldsymbol{h}_{Arg1})^T \boldsymbol{M}^{[1:k]} \boldsymbol{h}_{Arg2}$ results in a vector $\boldsymbol{m} \in \mathbb{R}^k$, where each entry is computed by one slice $i = 1, 2, ..., k$ of the tensor: $\boldsymbol{m}_i = (\boldsymbol{h}_{Arg1})^T \boldsymbol{M}^{[i]} \boldsymbol{h}_{Arg2}$. The other parameters $\boldsymbol{V} \in \mathbb{R}^{k \times 2d_h}$, $\boldsymbol{u} \in \mathbb{R}^k$, $\boldsymbol{b} \in \mathbb{R}^k$ are the standard form of a neural network. Here, each tensor slice can be seen as a "feature extractor", which extracts the features expressing the *Arg*1-*Arg*2 interactions.

Through the tensor layer, we can obtain the semantic interactions between two arguments as features, which are further reshaped to a vector and fed to a full connection hidden layer. Then we apply a softmax function in the output layer to compute the probabilities of different relations and recognize them.

2.2.2 Model Training with Sparse Constraint

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

$$L(\hat{\boldsymbol{y}}, \boldsymbol{y}) = -\sum_{j=1}^{C} \boldsymbol{y}_j \log(Pr(\hat{\boldsymbol{y}}_j)). \tag{13}$$

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

Based on argument representations with interactive attention, tensor embodies the different aspects of semantic interactions between two arguments. However, not all the interactions are useful. There could exist some redundant and noisy interactions influencing the system performance. In order to remove the irrelevant interactions and select the indicative pair patterns, the large portion of $\boldsymbol{M}^{[i]}$ should be zero. Therefore, we introduce the 1-norm regularizer to promote the feature sparsity. This element-wise sparsity can be helpful when most of the features are irrelevant to the learning objective. Furthermore, we also add l_2 regularization to avoid over-fitting issue. And the training objective function is transformed as:

$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{r=1}^{n} L(\hat{\boldsymbol{y}}^{(r)}, \boldsymbol{y}^{(r)}) + R(\boldsymbol{\theta}), \tag{14}$$

$$R(\boldsymbol{\theta}) = \lambda_M \|\boldsymbol{\theta}_M\|_1 + \frac{\lambda_O}{2}\|\boldsymbol{\theta}_O\|_2. \tag{15}$$

where $R(\boldsymbol{\theta})$ is the regularization term with respect to $\boldsymbol{\theta}$. We divide $\boldsymbol{\theta}$ into two parts: $\boldsymbol{\theta}_M$ is the tensor term weights, and $\boldsymbol{\theta}_O$ is the other parameters of our model. Especially, $\|\boldsymbol{\theta}_M\|_1$ in Eq.(15) is l_1 regularization for the tensor slices, which is used to filter the important values.

To minimize the objective, we employ the proximal gradient descent method (Parikh and Boyd, 2014) since l_1 regularization is non-differentiable at zero. It is used for optimizing the objective as a combination of both smooth and non-smooth terms. The update formulas is as follows:

$$\boldsymbol{\theta}_i^{(t')} = \boldsymbol{\theta}_i^{(t)} - \gamma\left(\frac{\partial L}{\partial \boldsymbol{\theta}_i} + \lambda\frac{\partial R}{\partial \boldsymbol{\theta}_i}\right)\Big|_{\boldsymbol{\theta}_i = \boldsymbol{\theta}_i^{(t)}}, \tag{16}$$

$$\frac{\partial R}{\partial \boldsymbol{\theta}_i} = \begin{cases} 2\boldsymbol{\theta}_i^{(t)}, & \text{if } \|\boldsymbol{\theta}\|_2; \\ sign(\boldsymbol{\theta}_i^{(t)}), & \text{if } \|\boldsymbol{\theta}\|_1, \text{ and } \boldsymbol{\theta}_i^{(t)} \neq 0. \end{cases} \tag{17}$$

$$\boldsymbol{\theta}_i^{(t+1)} = prox_{\lambda}(\boldsymbol{\theta}_i^{t'}) = \tau(\boldsymbol{\theta}_i^{t'}, \gamma\lambda), \tag{18}$$

$$\tau(a, z) = \begin{cases} a - z, & \text{if } a > z; \\ a + z, & \text{if } a < -z; \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

where $prox_{\lambda}$ is a proximal operator, τ is a soft-thresholding operator, and γ is the learning rate.

3 Experiments

3.1 Data Preparation

Corpus. We use the Penn Discourse TreeBank (PDTB) (Prasad et al., 2008), which is the largest hand-annotated discourse relation corpus annotated on 2312 Wall Street Journal (WSJ) articles. Experiments are conducted on the four top-level classes as in previous work (Rutherford and Xue, 2014; Chen et al., 2016). Following the conventional data splitting, we use Section 2-20 as training set, Section 21-22 as testing set, and Section 0-1 as development set. The relevant statistics is shown in Table 1.

Relation	**Train**	**Dev.**	**Test**
Comparison	1842	393	144
Contingency	3139	610	266
Expansion	6658	1231	537
Temporal	579	83	55

Table 1: Statistics of implicit discourse relations.

Hyper-parameters	**Value**
Initial learning rate	0.01
Minibatch size	30
Dropout rate	0.1
Number of tensor slice	3

Table 2: Hyper-parameters for our TIASL model.

Experimental Settings. The 50-dimensional pre-trained word embeddings are provided by GloVe (Pennington et al., 2014), which are fixed during our model training. All the discourse arguments are padded to the same length of 50. And the length of intermediate representation for our network is also 50. The other parameters are initialized by random sampling from uniform distribution in [-0.1,0.1]. We do not present the details of tuning the hyper-parameters and only give their final settings as shown in Table 2.

To evaluate our model, we adopt two kinds of experiment settings, including a four-way classification and four separate one-vs-other binary classification. The former is to observe the overall performance. And the latter is to solve the problem of unbalance data, where each top level class is against the other three discourse relation classes. We use an equal number of positive and negative instances in the training set in each class. The testing set and development set keep the natural state.

3.2 Comparison Methods

We choose the following models as our baselines, which are the state-of-the-art models in argument representation, interaction and attention aspects during implicit discourse relation recognition.

- **Ji2015**: Ji and Eisenstein (2014) utilized two recursive neural networks on the syntactic parse tree to induce argument representation and entity spans.
- **Qin2016**: Qin et al. (2016b) integrated a CNN and a Collaborative Gated Neural Network (CGNN) into argument representation.
- **Chen2016**: Chen et al. (2016) used a Gated Relevance Network (GRN) and incorporated both the linear and non-linear interactions between word pairs.
- **Liu2016**: Liu and Li (2016) designed Neural Networks with Multi-level Attention (NNMA) and selected the important words for recognizing discourse relation. Here, we select the models with two and three levels attention as baselines.

Besides, we also use the following variants of RNN and the proposed TIASL model for comparisons.

- **LSTM**: encode two discourse arguments by LSTM respectively, and concatenate the two representations, feeding them to the full connection hidden layer as the input of softmax classifier.
- **Bi-LSTM**: based on **LSTM**, we consider the bidirectional context information, and use Bi-LSTM to encode two discourse arguments.
- **Bi-LSTM+Interactive Attention**: further integrate the interactive attention to obtain the new argument representations shown in Eq.(11).
- **Bi-LSTM+Tensor Layer**: based on **Bi-LSTM**, adopt the neural tensor network to capture the semantic interaction between two arguments.
- **TIA with k-Max Pooling**: use k-max pooling operation instead of our sparse strategy to select features after tensor layer in neural Tensor network with Interactive Attention model (TIA).
- **Our TIASL**: based on **Bi-LSTM** with **Interactive Attention** and **Tensor Layer**, we add l_1 regularization for tensor parameters in order to capture the most important interactions.

3.3 The Overall Performance

Table 3 shows the overall performance, using F_1 score and accuracy for four-wary classification and F_1 score for binary classification. With respect to four-way classification, we have the following observations:

- Ji2015 gains the lowest performance on both F_1 score and accuracy, which separately computes discourse argument representations by integrating syntactic parse tree into RNN. It indicates a simple neural network, which ignores the interactive context of two discourse arguments, is not sufficient for implicit discourse relation recognition.

Model	Binary Classification				Four-way Classification	
	Comp.	Cont.	Expa.	Temp.	F1	Acc.
Ji2015	35.93%	52.78%	-	27.63%	38.52%	43.56%
Qin2016	41.55%	57.32%	71.50%	35.43%	-	-
Chen2016	40.17%	54.76%	-	31.32%	44.61%	57.84%
Liu2016 (two levels)	36.70%	54.48%	70.43%	38.84%	46.29%	57.17%
Liu2016 (three levels)	39.86%	53.69%	69.71%	37.61%	44.95%	57.57%
Our TIASL	40.35%	56.81%	72.11%	38.65%	47.59%	59.06%

Table 3: Comparisons with the state-of-the-art models.

- The accuracy of Chen2016 model is better than that of other baselines. It verifies the effectiveness word pair information, which uses gate mechanism to control the combination of linear and non-linear interactions between argument pairs. However, there unavoidably exits some noises, and this model has not considered the sparsity of pair patterns. F_1 score of Liu2016 (two levels) model is higher than that of other baselines, which achieves 1.34% than that of three levels attention's. It indicates that attention mechanism is useful, and yet paying more attention may bring the over-fitting problem due to more parameters.

- Our TIASL gains an improvement 1.30% on F_1 score than that of Liu2016 (two levels), and an improvement 1.22% on accuracy than that of Chen2016. The results imply that our model with interactive attention for the bidirectional asymmetry of two arguments and sparse pair pattern selection is useful for recognizing implicit discourse relation.

For binary classification, the observations are as follows:

- F_1 scores of Temporal relation are the lowest in all models. This is reasonable since it accounts for the smallest number of instances (only 5%) in the corpus. With the increase of instance number in different relations, F_1 scores also rise. It proves that the corpus is also crucial to implicit discourse relation recognition.

- Qin2016 gains the best performance on Contingency, and our TIASL model obtains the comparable score with it. Notably, Chen2016 and Liu2016 (two levels) are quite relevant work to ours. Different from them, our model integrates the attention-based interactive information between arguments at representation stage. This may be the main reason why our TIASL model is better than the two models (the improvements of 2.05% and 2.33%, respectively). Similar results are in Comparison relation.

- Our TIASL model achieves state-of-the-art performance in recognition of the Expansion relation. The reasons are two-fold: (1) some argument pairs may have confusable word pairs, which can be effectively mined by asymmetric attention; (2) some complex argument pairs need to be further understood their semantic representation and explore the indicative and interactive patterns. Our TIASL model integrates these two aspects and performs well.

3.4 The Effectiveness of Each Component

In order to verify the effectiveness of attention mechanism, neural tensor layer and l_1 regularization, we design five experiments to compare with our TIASL model. Seen from Table 4, we have the following observations:

- The performance of LSTM is the worst on each relation. Although Bi-LSTM captures more information than LSTM, the results are not very good. The reason is that separately encoding discourse argument by LSTM or Bi-LSTM ignores the local focused words since it equally treats every word.

Model	Binary Classification				Four-way Classification	
	Comp.	Cont.	Expa.	Temp.	F1	Acc.
LSTM	32.95%	43.38%	68.10%	30.80%	36.40%	54.50%
Bi-LSTM	34.01%	44.68%	68.53%	31.27%	36.54%	55.31%
Bi-LSTM + Interactive Attention	35.43%	45.92%	68.57%	32.50%	43.27%	55.68%
Bi-LSTM + Tensor Layer	37.36%	46.73%	69.81%	33.89%	43.61%	55.97%
TIA with k-max pooling	39.78%	55.13%	70.96%	37.65%	46.77%	57.83%
Our TIASL	40.35%	56.81%	72.11%	38.65%	47.59%	59.06%

Table 4: The effects of different components.

- Bi-LSTM with Interactive Attention performs better than the above two simple models. In detail, the F_1 score of this model gains 2.48%, 2.54%, 1.70% improvement on Comparison, Contingency and Temporal than that of LSTM, respectively. We perform significance test for these improvements, and they are both significant under one-tailed t-test ($p < 0.05$). It indicates that the model could find pair specific clues in two arguments by constructing the relevance of word pairs to some extent. And the effectiveness of our attention mechanism for capturing the interactive information between the arguments is crucial at representation stage.
- Bi-LSTM with Tensor Layer slightly achieves better performance. This indicates the effectiveness of tensor layer for capturing complex interactive features. The TIA model, which combines attention mechanism and tensor layer, also performs better, but lower than our TIASL model. It is because that k-max pooling strategy could not guarantee getting the important interaction pairs from the global perspective. How to represent and explain these interactive features mined in our model will be our next research focus.
- Our TIASL model achieves the best performance. It not only encodes discourse arguments with important word pairs by interactive attention, but also captures the more deeper and the more important semantic interactions by NTN with l_1 regularization. The integration of all components is useful for recognizing implicit discourse relations.

The observations of each component's four-way classification are consistent with the binary classification.

3.5 Interactive Attention Analysis

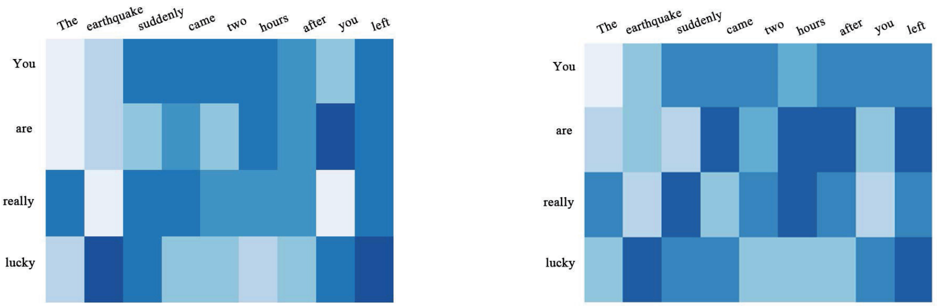
To demonstrate the validity of our interactive attention, we visualize the heat maps of argument pairs shown in Figure 2, which shows the interaction matrices of only using Bi-LSTM and our interactive attention on an example. Every word accompanies with the various background colors. The darker patches denote the correlations of word pairs are higher. The example of *Contingency* relation is listed below:

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

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

However, it might be classified as a *Comparison* relation if we only focus on the informative word pair (lucky, earthquake) with contrasting sentiment polarity. Therefore, we need to consider the context of the whole argument pair to infer the correct relation from two back and forth reading directions.

Seen from Figure 2(a), the word pairs (are, you), (lucky, earthquake), (lucky, left) get the higher scores, the scores on the other pairs are arbitrary. It demonstrates the Bi-LSTM model may be influenced by the word pair frequency in corpus. Meanwhile, it encodes two arguments separately, which ignores the relevant and informative interactions between two arguments. Figure 2(b) as a comparison, we observe that there are the more word pairs obtaining the higher scores, which are ignored in Figure 2(a). This proves that the effectiveness of generating interactive argument representation by our interactive attention, which imitates human-like reading strategy.



(a) Bi-LSTM (only word embedding)

(b) Bi-LSTM + Interactive attention

Figure 2: The relatedness between two arguments.

4 Related Work

Traditional methods for implicit discourse relation recognition rely on artificial and shallow features, such as POS, polarity, word position, etc. Recent neural network based methods acquire the better performance, and mainly focus on two aspects:

4.1 Argument Representation

The prerequisite of recognizing discourse relation is to have a good argument representation. Most previous researches use various neural networks, such as CNN, RNN, and hybrid models (Zhang et al., 2015; Qin et al., 2016a; Rutherford et al., 2016) to encode discourse arguments as low-dimensional, dense and continuous representations. Ji and Eisenstein (2014) integrate the linguistic features, including syntactic parsing and coreferent entity mentions into compositional distributed representations.

Though argument representation contains the high-level semantic, it does not embody emphasis during reading comprehension. Some used neural architectures with attention mechanism pick up the important information from discourse arguments (Mnih et al., 2014; Zhang et al., 2016). Li et al. (2016) exploit the hierarchical attention to capture the focus of different granularity. Liu and Li (2016) imitate the repeated reading strategy, and proposes neural networks with multi-level attention to recognize discourse relations. However, these researches have not considered the human-like reading strategy from two directions. The imagination by first reading one argument is different from the other, which has the reciprocal effects on implicit discourse relation recognition.

4.2 Pair Interactions

The emphasis of discourse arguments is partly obtained by attention mechanism. Most studies tend to discover more semantic interactions between two arguments by complex neural networks (Chen et al., 2016; Qin et al., 2016b; Lan et al., 2017). Cai and Zhao (2017) generate discourse argument representations via pair-specified feature extraction. Lei et al. (2017) conduct word interaction score to capture both linear and quadratic relation for argument representation.

Neural tensor network is good at capturing multiple interactions in data, and gets the good performance on entity relation (Socher et al., 2013a), Chinese word segmentation(Pei et al., 2014) and sentiment analysis (Socher et al., 2013b) tasks. And some NTN-like methods learn the semantic interaction between discourse arguments (Chen et al., 2016). Yet they do not discriminate the noises and the redundant information existed in interactions, and ignore the sparsity of pair patterns.

Inspired by sparse learning in deep neural networks (Collins and Kohli, 2014; Yoon and Hwang, 2017; Wen et al., 2017), they use sparse regularization to obtain compact deep networks by removing unnecessary weighs. In our paper, we introduce sparse learning into neural tensor network to select some indicative and informative word pair patterns. To our knowledge, our study is the first to employ the idea of sparse learning in implicit discourse relation recognition.

5 Conclusion

A novel neural tensor network framework with interactive attention and sparse learning (TIASL) is proposed for implicit discourse relation recognition. We imitate human-like bidirectional reading strategy, and encode the semantic representation with reciprocal influence of discourse arguments through interactive attention. And we further adopt neural tensor network with l_1 regularization to capture the indicative and informative interactions between discourse arguments. Our experimental results on PDTB show that the proposed TIASL model is effective.

However, we just take the surface word pairs to express the correlation by calculating the pair-wise matrix in this paper. We will automatically mine the deeper interaction between two arguments and explain the specific patterns of the different relations.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (61472277, 61772361, 61771333). We also thank the anonymous reviewers for their valuable comments.

References

Chloé Braud and Pascal Denis. 2015. Comparing word representations for implicit discourse relation classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2201–2211.

Deng Cai and Hai Zhao. 2017. Pair-aware neural sentence modeling for implicit discourse relation classification. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 458–466. Springer.

Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. 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 (ACL)*, pages 1726–1735.

Maxwell D Collins and Pushmeet Kohli. 2014. Memory bounded deep convolutional networks. *arXiv preprint arXiv:1412.1442*.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning (ICML)*, pages 160–167. ACM.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1602–1613.

Herzog, Kammer Michael H., and Scharnowski Frank Thomas. 2016. Time slices: What is the duration of a percept? *PLOS Biology*.

Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. 2014. Towards an open-domain conversational system fully based on natural language processing. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 928–939.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Rodolphe Jenatton, Nicolas L Roux, Antoine Bordes, and Guillaume R Obozinski. 2012. A latent factor model for highly multi-relational data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3167–3175.

Yangfeng Ji and Jacob Eisenstein. 2014. One vector is not enough: Entity-augmented distributional semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, 3:329–344.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. 2017. 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 (EMNLP)*, pages 1299–1308.

Wenqiang Lei, Xuancong Wang, Meichun Liu, Ilija Ilievski, Xiangnan He, and Min-Yen Kan. 2017. Swim: A simple word interaction model for implicit discourse relation recognition. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4026–4032.

Qi Li, Tianshi Li, and Baobao Chang. 2016. Discourse parsing with attention-based hierarchical neural networks. In *Proceedings of the 2016 Conference on Emirical Methods in Natural Language Processing (EMNLP)*, pages 362–371.

Yang Liu and Sujian Li. 2016. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1224–1233.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, pages 2750–2756.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. 2014. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2204–2212.

Neal Parikh and Stephen Boyd. 2014. Proximal algorithms. *Foundations and Trends in Optimization.*, 1(3):127–239.

Wenzhe Pei, Tao Ge, and Baobao Chang. 2014. Max-margin tensor neural network for chinese word segmentation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 293–303.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind K Joshi. 2008. Easily identifiable discourse relations. *Technical Reports (CIS)*.

Emily Pitler, Annie Louis, and Ani and Nenkova. 2009. 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*, pages 683–691.

Rashmi Prasad, Nikhil Diesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC)*.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016a. Shallow discourse parsing using convolutional neural network. In *CoNLL Shared Task*, pages 70–77.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016b. A stacking gated neural architecture for implicit discourse relation classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2263–2270.

Attapol Rutherford and Nianwen Xue. 2014. 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 (EACL)*, pages 645–654.

Attapol T Rutherford, Vera Demberg, and Nianwen Xue. 2016. Neural network models for implicit discourse relation classification in english and chinese without surface features. *arXiv preprint arXiv:1606.01990*.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013a. Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems*, pages 926–934.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013b. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642.

Wei Wen, Yuxiong He, Samyam Rajbhandari, Wenhan Wang, Fang Liu, Bin Hu, Yiran Chen, and Hai Li. 2017. Learning intrinsic sparse structures within long short-term memory. *arXiv preprint arXiv:1709.05027*.

Jaehong Yoon and Sung Ju Hwang. 2017. Combined group and exclusive sparsity for deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning (PMLR)*, pages 3958–3966.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow convolutional neural network for implicit discourse relation recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2230–2235.

Biao Zhang, Deyi Xiong, and Jinsong Su. 2016. Neural discourse relation recognition with semantic memory. *arXiv preprint arXiv:1603.03873*.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1507–1514. Association for Computational Linguistics.