Multi-task Attention-based Neural Networks for Implicit Discourse Relationship Representation and Identification

Man Lan[1,2], **Jianxiang Wang**[1,3*], **Yuanbin Wu**[1,2*], **Zheng-Yu Niu**[3*] , **Haifeng Wang**[3]
[1] School of Computer Science and Software Engineering, East China Normal University
[2] Shanghai Key Laboratory of Multidimensional Information Processing, P.R.China
[3] Baidu Inc., Beijing, P.R.China
{mlan,ybwu}@cs.ecnu.edu.cn
{wangjianxiang01,niuzhengyu,wanghaifeng}@baidu.com

Abstract

We present a novel multi-task attention-based neural network model to address implicit discourse relationship representation and identification through two types of representation learning, an attention-based neural network for learning discourse relationship representation with two arguments and a multi-task framework for learning knowledge from annotated and unannotated corpora. The extensive experiments have been performed on two benchmark corpora (i.e., PDTB and CoNLL-2016 datasets). Experimental results show that our proposed model outperforms the state-of-the-art systems on benchmark corpora.

1 Introduction

The task of implicit discourse relation (or rhetorical relation) identification is to recognize how two adjacent text spans without explicit discourse marker (i.e., connective, e.g., *because* or *but*) between them are logically connected to one another (e.g., *cause* or *contrast*). It is considered to be a crucial step for discourse analysis and language generation and helpful to many downstream NLP applications, e.g., QA, MT, sentiment analysis, machine comprehension, etc.

With the release of PDTB 2.0 (Prasad et al., 2008), lots of work has been done for discourse relation identification on natural (i.e., genuine) discourse data (Pitler et al., 2009; Lin et al., 2009; Wang et al., 2010; Zhou et al., 2010; Braud and Denis, 2015; Fisher and Simmons, 2015) with the use of traditional NLP linguistically informed features and machine learning algorithms. Recently, more and more researchers resorted to neural networks for implicit discourse recognition (Zhang et al., 2015; Chen et al., 2016; Liu et al., 2016b; Qin et al., 2016a; Liu and Li, 2016; Braud and Denis, 2016; Wu et al., 2016). Meanwhile, to alleviate the shortage of labeled data, researchers explored multi-task learning with the aid of unannotated data for implicit discourse recognition either in traditional machine learning framework (Collobert and Weston, 2008; Lan et al., 2013) or recently in neural network framework (Wu et al., 2016; Liu et al., 2016b).

In this work, we present a novel multi-task attention-based neural network to address implicit discourse relationship representation and recognition. It performs two types of representation learning at the same time. An attention-based neural network conducts discourse relationship representation learning through interaction between two discourse arguments. Meanwhile, a multi-task learning framework leverages knowledge from auxiliary task to enhance the performance of main task. Furthermore, these two types of learning are integrated into one neural network framework and work together to maximize the overall performance.

The contributions of this work are listed as follows.

- We propose a multi-task attention-based neural network model to address implicit discourse relationship representation and recognition, which benefits from both the interaction between discourse arguments and the interaction between different learning tasks;
- Our method achieves the best results on two benchmark corpora in comparison with the state-of-the-art systems so far.

The organization of this work is as follows. Section 2 describes the proposed novel multi-task neural network. Section 3 introduces the exper-

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1308–1317
Copenhagen, Denmark, September 7–11, 2017. ©2017 Association for Computational Linguistics

imental settings in detail. Section 4 reports the comprehensive experimental results on two benchmark corpora. Section 5 summarized related work. Finally, Section 6 concludes this work.

2 Multi-task Attention-based Neural Networks Models

2.1 Motivation

The idea of learning two types of interactive knowledge from arguments and from multi-tasks is motivated by the following observations and analysis.

On the one hand, to recognize the discourse relationships, our system needs to understand the meaning of each argument and infer the discourse sense transferred between two arguments (denoted as *Arg*-1 and *Arg*-2). Learning the semantic representation of each argument (sentence) has been studied with the use of many neural network models and their variants (e.g., CNN, RNN, LSTM, Bi-LSTM, ect). However, learning the complicated and various types of discourse relationships between arguments may not be performed by simply summing up or concatenating two argument representations. We analyze the discourse with contrast relationship and find that the contrast information may result from different parts of sentence, e.g., tenses (e.g., previous vs. now), entities (their vs our), or even the whole arguments, etc. Therefore, in order to learn the relationship representation between two arguments, we propose an attention mechanism that can select out the most important part from two arguments and perform the information interaction between two arguments.

On the other hand, one common issue involved in implicit discourse relationship identification is the lack of labeled data. In this work, we state that the relevant information from unlabelled data might be helpful and we present a novel multi-task learning framework. In contrast with previous multi-task learning framework in traditional machine learning, we improve multi-task learning framework with representation learning for better discourse relationship representation.

Inspired by the above considerations, we present a novel multi-task attention-based neural network model by integrating attention mechanism with multi-task learning for information interaction between arguments and between tasks.

2.2 Learning Discourse Representation

To learn the semantic representation of each argument in discourse, a lot of neural network models and their variants have been proposed, such as, convolutional neural network (CNN), recurrent neural network (RNN) and so on. As a variant of RNN, long-short term memory (LSTM) neural network specifically addresses the issue of learning long-term dependencies and is good at modeling over a sequence of words with consideration of the contextual information. Therefore, in this work we adopt LSTM to model discourse argument.

2.2.1 LSTM for Argument Representation

Figure 1 shows the traditional LSTM model for representation learning of arguments. First of all, through the embedding layer, we associate each word w in the vocabulary with a vector representation $\boldsymbol{x}_w \in \mathbb{R}^{d_w}$. Let $\boldsymbol{x}_i^1$ ($\boldsymbol{x}_i^2$) be the i-th word vector in *Arg*-1 (*Arg*-2), then these two discourse arguments are represented as:

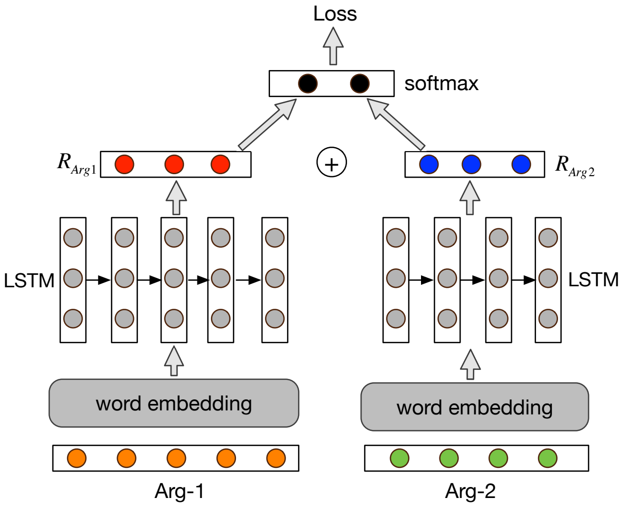

Figure 1: LSTM for discourse argument pair representation learning.

$$Arg\text{-}1\text{: } [\boldsymbol{x}_1^1, \boldsymbol{x}_2^1, \cdots, \boldsymbol{x}_{L_1}^1] \quad (1)$$

$$Arg\text{-}2\text{: } [\boldsymbol{x}_1^2, \boldsymbol{x}_2^2, \cdots, \boldsymbol{x}_{L_2}^2] \quad (2)$$

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

Given the word representations of the argument $[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_L]$ as the input sequence, an LSTM computes the state sequence $[\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_L]$ for each time step i using the following formulation:

$$\boldsymbol{i}_i = \sigma(\boldsymbol{W}_i[\boldsymbol{x}_i, \boldsymbol{h}_{i-1}] + \boldsymbol{b}_i) \quad (3)$$

$$\boldsymbol{f}_i = \sigma(\boldsymbol{W}_f[\boldsymbol{x}_i, \boldsymbol{h}_{i-1}] + \boldsymbol{b}_f) \quad (4)$$

$$\boldsymbol{o}_i = \sigma(\boldsymbol{W}_o[\boldsymbol{x}_i, \boldsymbol{h}_{i-1}] + \boldsymbol{b}_o) \quad (5)$$

$$\tilde{\boldsymbol{c}}_i = tanh(\boldsymbol{W}_c[\boldsymbol{x}_i, \boldsymbol{h}_{i-1}] + \boldsymbol{b}_c) \quad (6)$$

$$\boldsymbol{c}_i = \boldsymbol{i}_i \odot \tilde{\boldsymbol{c}}_i + \boldsymbol{f}_i \odot \boldsymbol{c}_{i-1} \quad (7)$$

$$\boldsymbol{h}_i = \boldsymbol{o}_i \odot tanh(\boldsymbol{c}_i) \quad (8)$$

where [ ] means the concatenation operation of vectors, σ denotes the sigmoid function and $\odot$ denotes element-wise product. Besides, $\boldsymbol{i}_i$, $\boldsymbol{f}_i$, $\boldsymbol{o}_i$ and $\boldsymbol{c}_i$ denote the input gate, forget gate, output gate and memory cell, respectively. Moreover, we also use bidirectional LSTM (Bi-LSTM) which is able to capture the context from both past and future rather than LSTM which only considers the context information from the past. Therefore, 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 sequence states $[\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_L]$ to get the representations of *Arg*-1 and *Arg*-2 respectively as follows:

$$\boldsymbol{R}_{Arg_1} = \sum_{i=1}^{L_1} \boldsymbol{h}_i^1 \quad (9)$$

$$\boldsymbol{R}_{Arg_2} = \sum_{i=1}^{L_2} \boldsymbol{h}_i^2 \quad (10)$$

Finally we concatenate the two argument representations $\boldsymbol{R}_{Arg_1}$ and $\boldsymbol{R}_{Arg_2}$ as the argument pair representation, i.e., $\boldsymbol{R}_{pair} = [\boldsymbol{R}_{Arg_1}, \boldsymbol{R}_{Arg_2}]$.

Clearly, in this way, there is no any correlation and interaction between the two arguments. That is, whatever the types of discourse relationship they hold, the argument pair representation $\boldsymbol{R}_{pair}$ is independent from $\boldsymbol{R}_{Arg_1}$ or $\boldsymbol{R}_{Arg_2}$.

2.2.2 Attention Neural Network for Relationship Representation

In order to effectively capture the complicated and various types of relationships between arguments, we proposed a novel attention-based neural network model shown in Figure 2.

To do it, we first compute the match between $\boldsymbol{R}_{Arg_1}$ ($\boldsymbol{R}_{Arg_2}$) and each state $\boldsymbol{h}_i^2$ ($\boldsymbol{h}_i^1$) of *Arg*-2 (*Arg*-1) by taking the inner product followed by a softmax as follows:

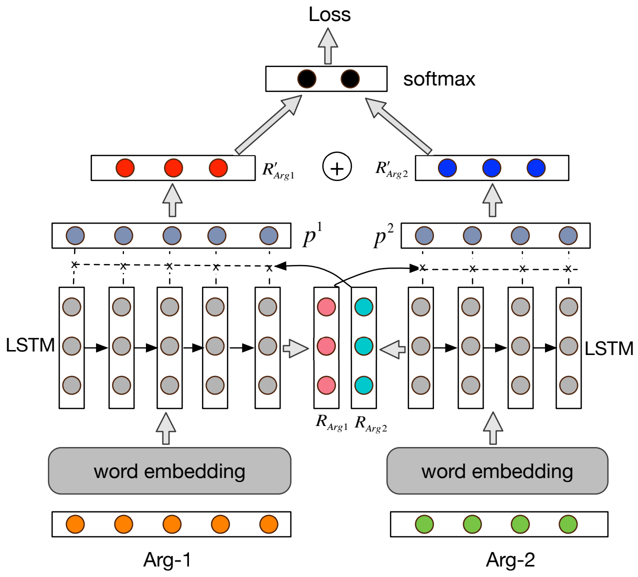

Figure 2: Attention Neural Network for representation learning of arguments.

$$p_i^1 = \text{Softmax}(\boldsymbol{R}_{Arg_2}^T \boldsymbol{h}_i^1) \quad (11)$$

$$p_i^2 = \text{Softmax}(\boldsymbol{R}_{Arg_1}^T \boldsymbol{h}_i^2) \quad (12)$$

where $\text{Softmax}(z_i) = e^{z_i} / \sum_j e^{z_i}$. Here $\boldsymbol{p}$ is an attention (probability) vector over the inputs and can be viewed as the weights of the words measuring to what degree our model should pay attention to. It is worth noting that $\boldsymbol{p^1}$ and $\boldsymbol{p^2}$ are determined by $\boldsymbol{R}_{Arg_2}$ and $\boldsymbol{R}_{Arg_1}$ respectively, which means the representation of one argument depends on the representation of the other.

Next, we sum over the state $\boldsymbol{h}_i$ weighted by the attention vector $\boldsymbol{p}$ to compute the new representations for *Arg*-1 and *Arg*-2 respectively as below:

$$\boldsymbol{R}'_{Arg_1} = \sum_{i=0}^{L_1} \boldsymbol{h}_i^1 p_i^1 \quad (13)$$

$$\boldsymbol{R}'_{Arg_2} = \sum_{i=0}^{L_2} \boldsymbol{h}_i^2 p_i^2 \quad (14)$$

The representation of *Arg*-2 ($\boldsymbol{R}_{Arg_2}$) is used to compute the weights of words in *Arg*-1 (i.e., $\boldsymbol{p^1}$) and $\boldsymbol{R}_{Arg_1}$ is used to compute the weights of words in *Arg*-2 (i.e., $\boldsymbol{p^2}$). In this way, the new representations of the two arguments interact with each other. Therefore, this attention mechanism enables our model to focus on specific spans in the two arguments, which is crucial to recognize the discourse relations. We then concatenate $\boldsymbol{R}'_{Arg_1}$

and $\boldsymbol{R}'_{Arg_2}$ to get the argument pair representation $\boldsymbol{R}_{pair} = [\boldsymbol{R}'_{Arg_1}, \boldsymbol{R}'_{Arg_2}]$.

Finally, we feed the argument pair vector $\boldsymbol{R}_{pair}$ to a fully-connected softmax layer which outputs the probabilities of different classes for the classification task. Here we choose the cross-entropy loss between the outputs of the softmax layer and the ground-truth class labels as our loss function.

2.3 Multi-task Attention-based Neural Networks

The model presented in Section 2.2 can perform implicit discourse relation recognition in itself. However, similar with many models in deep learning, one big issue is the lack of labeled data. Therefore, we propose a multi-task attention-based neural network by integrating the aforementioned model into a multi-task learning framework to address the implicit discourse relation recognition with the aid of large amount of unlabelled data. Figure 3 shows the general framework of our proposed multi-task attention-based neural network model.

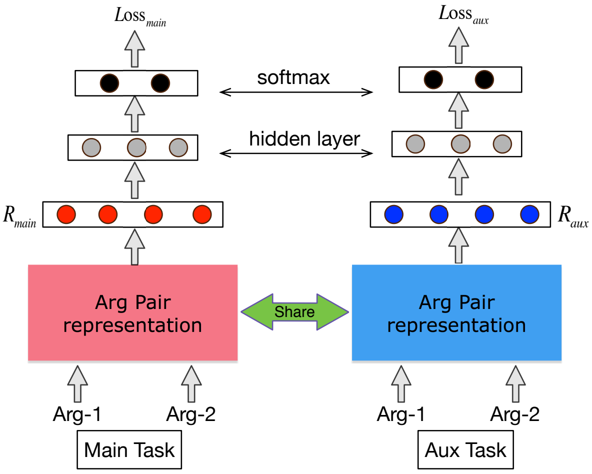

Figure 3: The framework of our proposed multi-task attention-based neural network model.

We use the aforementioned attention-based neural network to map the argument pair into a low-dimensional vector ($\boldsymbol{R}_{pair}$) denoted as `Arg Pair representation` component in Figure 3. Under the multi-task learning framework, the parameters of the `Arg Pair representation` components are shared between the main task and the auxiliary tasks. We denote $\boldsymbol{R}_{main}$ and $\boldsymbol{R}_{aux}$ as the representations of argument pair for main and auxiliary tasks, respectively. And we add a hidden layer after $\boldsymbol{R}_{main}$ and $\boldsymbol{R}_{aux}$ to learn the task-specific representations followed by the softmax layers used to compute the loss of the main task ($Loss_{main}$) and the loss of the auxiliary task ($Loss_{aux}$), respectively.

Regarding the strategy of sharing knowledge learnt from auxiliary to main task, we propose the following three methods.

2.3.1 Equal Share

A simple and straightforward way is to equally share the knowledge learned from main task and auxiliary task. Therefore, the total loss of the multi-task neural network is calculated as:

$$Loss = Loss_{main} + Loss_{aux} \tag{15}$$

where $Loss_{aux}$ has the same weight as $Loss_{main}$.

2.3.2 Weighted Share

Another method is to give different weights to the main and auxiliary task as below:

$$Loss = Loss_{main} + w * Loss_{aux} \tag{16}$$

where $w \in (0, 1]$ is a weight parameter. Clearly, a lower value of w means less importance of auxiliary task.

2.3.3 Sigmoid (Gated) Interaction

The above two ways of sharing knowledge actually have no deep interaction between the main and auxiliary tasks. They only share equal or weighted contributions from tasks to final result. Therefore, we propose a model that can perform interaction between tasks, which is shown in Figure 4.

We introduce two important parameters $\boldsymbol{W}_{inter} \in \mathbb{R}^{d_{pair} \times d_{pair}}$ and $\boldsymbol{b}_{inter} \in \mathbb{R}^{d_{pair}}$ (d_{pair} is the length of the argument pair representation vector $\boldsymbol{R}_{pair}$) to fulfil the interaction between main and auxiliary tasks. As shown in the following Formula (17) and (18), the new representation of argument pair $\boldsymbol{R}'_{main}$ is updated by the combination of $\boldsymbol{W}_{inter}$ and $\boldsymbol{R}_{aux}$ using a Sigmoid function.

$$\boldsymbol{R}'_{main} = \boldsymbol{R}_{main} \odot \sigma(\boldsymbol{W}_{inter}\boldsymbol{R}_{aux} + \boldsymbol{b}_{inter}) \tag{17}$$

$$\boldsymbol{R}'_{aux} = \boldsymbol{R}_{aux} \odot \sigma(\boldsymbol{W}_{inter}\boldsymbol{R}_{main} + \boldsymbol{b}_{inter}) \tag{18}$$

$\boldsymbol{W}_{inter}$ and the Sigmoid function (σ) work together to make information interacted between two tasks and select useful relevant information out of the opposite tasks as well. Clearly, $\boldsymbol{W}_{inter}$ is

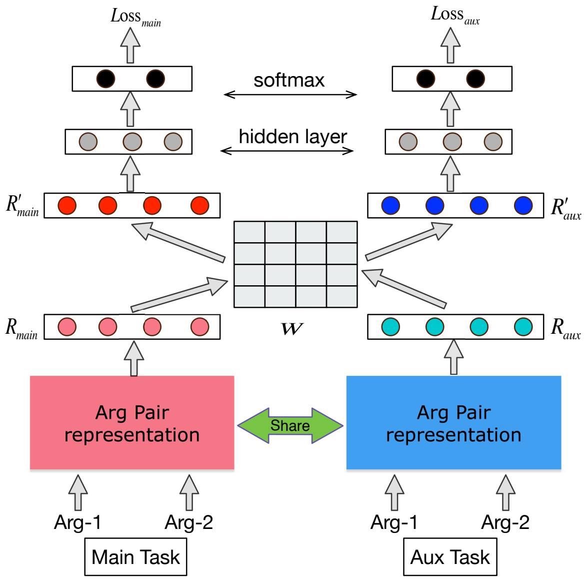

Figure 4: Sigmoid (Gated) interaction shared in multi-task framework (GShare).

a parameter to be trained. This mechanism acts as a gate to determine how much the information would pass through to the final result. Therefore, under the framework of multi-task and gated mechanism, the main and auxiliary tasks are capable of not only sharing the parameters of learning argument pair representation but also interacting the representations learning from each other.

2.4 Parameter Learning

We tried various settings of word embeddings trained on the BLLIP corpus with different dimensions d_{WE} = [50, 100, 150, 200] by *word2vec tool*[1] and finally set dimensionality as 50 based on the results on development set. we also explored the hidden state d_h = [50, 100, 150, 200] and the size of hidden layer in multi-task framework $d_{multi-task}$ = [50, 80, 120, 150]. Finally, for binary classification and four way classification on PDTB, we chose $d_h = 50$ and $d_{multi-task} = 80$. For multi-class classification on CoNLL-2016, we chose $d_h = 100$ and $d_{multi-task} = 120$. We applied dropout to the penultimate layer and set the dropout rate as 0.5. These parameters remain the same in experiments except the share weight w varies which will be discussed later. We chose the cross-entropy loss as loss function and adopted AdaGrad (Duchi et al., 2011) with a learning rate of 0.001 and a minibatch size of 64 to train the model.

3 Experiment Settings

3.1 Datasets

We adopted three corpora: PDTB 2.0 and CoNLL-2016 datasets are annotated for discourse relation recognition evaluation, and the BLLIP corpus is unlabeled and used as auxiliary task.

PDTB 2.0 is the largest annotated corpus of discourse relations, which contains 2, 312 Wall Street Journal (WSJ) articles. The sense label of discourse relations is hierarchically with three levels, i.e., class, type and sub-type. The top level contains four major semantic classes: Comparison (denoted as Comp.), Contingency (Cont.), Expansion (Exp.) and Temporal (Temp.). For each class, a set of types is used to refine relation sense. The set of subtypes is to further specify the semantic contribution of each argument. We focus on the top level (class) relations. Following (Pitler et al., 2009), we used sections 2-20 as training set, sections 21-22 as test set, and sections 0-1 as development set. Table 1 summarizes the statistics of four top level implicit discourse relations in PDTB.

Relation	Train	Dev	Test
Comp.	1942	197	152
Cont.	3342	295	279
Exp.	7004	671	574
Temp.	760	64	85

Table 1: The statistics of four top level implicit discourse relations in PDTB 2.0.

The CoNLL-2016 Shared Task focuses on shallow discourse parsing, which provides two test datasets, i.e., one from PDTB section 23 denoted as CoNLL-Test set, and the other from a similar source and domain (English Wikinews[2]) denoted as CoNLL-Blind test set. Different from the sense labels in PDTB, the CoNLL-Test set has three sense levels and the EntRel label. Moreover, it merges several labels in the original annotation to reduce some sparsity without losing too much of the utility and the semantics of the sense.

BLLIP The North American News Text (Complete) is used as unlabeled data source to generate synthetic labeled data for auxiliary task. We remove the explicit discourse connectives from raw texts and grant the explicit relations as the synthetic implicit relations. We obtain a resulting corpus with 100, 000 implicit relations by random sampling.

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

[2] https://en.wikinews.org/

3.2 Evaluation Measures

We adopt precision (*P*), recall (*R*) and their harmonic mean, i.e., F_1 for performance evaluation. We also report accuracy for direct comparison with previous works.

4 Results and Discussion

4.1 Results on PDTB in multiple binary classification

To be consistent with previous work, we first perform multiple binary classification (one-versus-other) on the four top level classes in PDTB. Several previous studies merged EntRel with Expansion, which is also explored in our study and noted as Exp+. Table 2 shows the results of our proposed three models in terms of F_1 (%) on PDTB using multiple binary classification, where STL means single task learning, *Eshare*, *Wshare* and *Gshare* denote the equal share, weighted share and gated interaction share under multi-task framework respectively, *Imp* denotes the standard implicit relations dataset in PDTB (similarly, *Imp* denotes standard implicit relations dataset in the CoNLL dataset when we perform experiments on the CoNLL dataset) used for training, *Exp* denotes all explicit relations in sections 00-24 in PDTB (similarly, all explicit relations in the CoNLL dataset when we perform experiments on the CoNLL dataset), and *BLLIP* denotes the synthetic implicit relations extracted from BLLIP. For example, *Imp* + *BLLIP* indicates that *Imp* is used for main task and *BLLIP* is for auxiliary task.

The first three rows in Table 2 list the results of LSTM, Bi-LSTM and attention neural network in the single task learning (STL) framework, which act as baselines for comparison with multi-task learning. We see that Bi-LSTM achieve slightly better performance than LSTM, which is consistent with previous work as Bi-LSTM considers the forward and backward direction contextual information while LSTM only considers the forward information. Compared with LSTM and Bi-LSTM, the attention neural network achieves much better performance. This indicates the effectiveness of attention mechanism for capturing the interaction between discourse arguments, which is crucial for relationship representation.

Generally, under the multi-task neural network framework, the three proposed multi-task neural networks, i.e., *Eshare*, *Wshare* and *Gshare*, outperform the single task learning methods. Comparing with *Eshare* and *Wshare*, we see that using a low value of w is able to boost the performance and reduce the negative influence brought by auxiliary task. We then use the best w value in *Wshare* to construct the loss of *Gshare* and the *Gshare* achieves the best performance among all methods through information interaction between main and auxiliary tasks.

Comparing *Imp* + *Exp* with *Imp* + *BLLIP*, we see that using *Exp* as auxiliary task achieves lower performance than using *BLLIP* and even hurts the performance compared with the single task. The possible reasons may result from (1) there is difference between explicit and implicit discourse relations and (2) the size of *Exp* dataset is much smaller than that of *BLLIP* and thus it is not large enough to boost the performance.

4.2 Results on PDTB and CoNLL-2016 in multi-class classification

We also perform multi-class classification on PDTB and CoNLL-2016. That is, a four-way classification on the four top-level classes in PDTB and a 15-way classification on the 15 sense labels in CoNLL dataset. Table 3 shows the results of multi-class classification on PDTB and CoNLL-2016 corpora in terms of accuracy (%) and macro-averaged F_1 (%).

The results of multi-class classification are consistent with the results of binary classification. First, the attention neural network achieves better performance than LSTM and Bi-LSTM. Second, the multi-task learning methods outperform the single-task learning method. Thrid, the *Gshare* method achieves the best performance.

4.3 Comparison with the state-of-the-art Systems

Table 4 lists the performance of our best model with the reported state-of-the-art systems on PDTB and CoNLL-2016. We see that our model achieves F_1 improvements of 1.64% on Cont., 0.97% on Exp., and 1.35% on Exp.+ against the best reported systems in binary classification. And in multi-class classification, our model also achieves the best performance of F_1 in four-way classification and accuracy in CoNLL-2016 Blind test set, which indicates that our model has good generality.

Specially, (Liu et al., 2016b) and (Liu and Li, 2016) listed in Table 4, which adopted neural network-based multi-task framework, are quite

		Comp.	Cont.	Exp.	Exp+	Temp
STL	LSTM	33.50	52.09	67.51	76.12	27.88
	Bi-LSTM	33.82	52.30	67.47	76.36	29.01
	Attention	**38.15**	**56.07**	**70.53**	**79.80**	**36.72**
Eshare	*Imp + Exp*	35.07	54.62	69.97	79.15	34.57
	Imp + BLLIP	37.67	56.82	70.81	80.43	35.48
Wshare	*Imp + Exp*	37.51 (w=0.1)	55.83 (w=0.2)	70.37 (w=0.3)	80.22(w=0.2)	35.71 (w=0.3)
	Imp + BLLIP	39.13 (w=0.2)	57.78(w=0.2)	71.88(w=0.1)	80.84 (w=0.3)	37.76(w=0.3)
Gshare	*Imp + Exp*	38.91	56.91	71.41	80.02	36.92
	Imp + BLLIP	**40.73**	**58.96**	**72.47**	**81.36**	**38.50**

Table 2: Performance of multiple binary classification on the top level classes in PDTB corpus in terms of F_1 (%).

		PDTB (Four way)	CoNLL-Test (*Acc*)	CoNLL-Blind (*Acc*)
STL	LSTM	F_1: 36.16; *Acc*: 56.12	34.45	35.07
	Bi-LSTM	F_1: 36.54; *Acc*: 54.30	34.85	35.83
	Attention	F_1: **45.57**; *Acc*: **57.55**	**37.41**	**38.36**
Eshare	*Imp + Exp*	F_1: 44.17; *Acc*: 55.65	35.56	37.06
	Imp + BLLIP	F_1: 44.57; *Acc*: 55.85	36.66	38.28
Wshare	*Imp + Exp*	F_1: 45.03; *Acc*: 56.21 (w=0.3)	36.24 (w=0.2)	37.34 (w=0.3)
	Imp + BLLIP	F_1: 45.80; *Acc*: **58.95** (w=0.2)	38.13 (w=0.1)	39.14 (w=0.4)
Gshare	*Imp + Exp*	F_1: 45.70; *Acc*: 57.17	37.84	38.10
	Imp + BLLIP	F_1: **47.80**; *Acc*: 57.39	**39.40**	**40.12**

Table 3: Performance of multi-class classification on PDTB and CoNLL-2016 in terms of accuracy (*Acc*) (%) and macro-averaged F_1 (%).

	Binary Classification (F_1)					Multi-class Classification (*Acc*)		
	Comp.	Cont.	Exp.	Exp+	Temp	PDTB (Four way)	CoNLL-Test(*Acc*)	CoNLL-Blind(*Acc*)
(Chen et al., 2016)	40.17	54.76	-	80.62	31.32	-	-	-
(Qin et al., 2016b)	**41.55**	57.32	71.50	80.96	35.43	-	-	-
(Liu and Li, 2016)	39.86	54.48	70.43	80.86	**38.84**	F_1: 46.29; *Acc*: **57.57**	-	-
(Wu et al., 2016)	-	-	-	-	-	F_1: 42.50; *Acc*: -	-	-
(Qin et al., 2016a)	38.67	54.91	-	80.66	32.76	-	-	-
(Liu et al., 2016b)	37.91	55.88	69.97	-	37.17	F_1: 44.98; *Acc*: 57.27	-	-
(Lan et al., 2013)	31.53	47.52	70.01	-	29.51	-	-	-
(Wang and Lan, 2016)	-	-	-	-	-	-	**40.91**	34.20
(Rutherford and Xue, 2016)	-	-	-	-	-	-	36.13	37.67
Our model	40.73	**58.96**	**72.47**	**81.36**	38.50	F_1: **47.80**; *Acc*: 57.39	39.40	**40.12**

Table 4: Comparison with the state-of-the-art systems reported on PDTB and CoNLL-2016, where - means N.A.

relevant to this work. (Liu et al., 2016b) presented a multi-task neural network, which considered information sharing between the main and auxiliary task. Different from their work, our work integrates the attention-based interaction between arguments and the multi-task based interaction between tasks into the final model. This is the main reason why our model achieves better performance in all types of relations, which shows the effectiveness of integrating gated mechanism into multi-task framework. Besides, (Liu and Li, 2016) used a complicated multi-level attention mechanism and the performance of our attention neural network in the single task is comparable to their results. Our multi-task attention model achieves better performance in most types with the aid of multi-task framework.

Besides, our previous work in (Lan et al., 2013) listed in Table 4, also presented a multi-task framework with traditional machine learning method to address implicit discourse recognition using BLLIP to obtain synthetic data. Clearly, under neural network-based multi-task framework, the attention and gated mechanism significantly improved the results and outperformed traditional machine learning method in all types of relations.

4.4 Effects of parameters w

Figure 5 shows the performance of four binary classification on four top level classes influenced by different share weights w in *Wshare* multi-task framework. We see that the best performance is achieved when we use a lower value of w. This

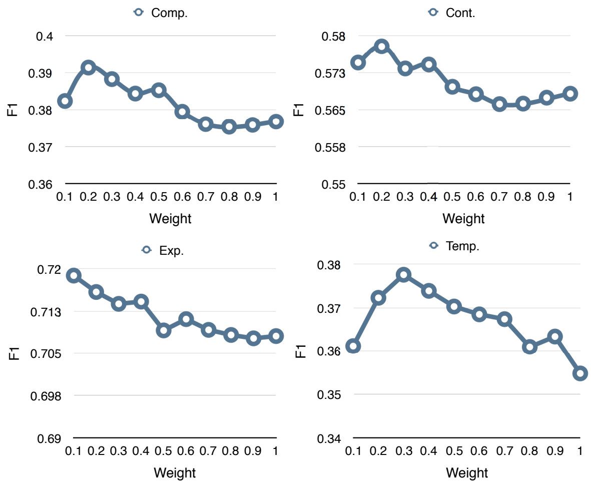

Figure 5: Results of top level implicit discourse relations in PDTB 2.0 with different weights w.

indicates that a low value of w can boost performance and reduce the negative influence brought by auxiliary task and enable our model to pay more attention to the main task.

5 Related Work

5.1 Implicit Discourse

With the release of PDTB 2.0, a number of studies performed discourse relation recognition on natural (i.e., genuine) discourse data with the use of traditional NLP techniques to extract linguistically informed features and traditional machine learning algorithms (Pitler et al., 2009; Lin et al., 2009; Wang et al., 2010; Braud and Denis, 2015; Fisher and Simmons, 2015).

Later, to make a full use of unlabelled data, several studies performed multi-task or unsupervised learning methods (Lan et al., 2013; Braud and Denis, 2015; Fisher and Simmons, 2015; Rutherford and Xue, 2015).

Recently, with the development of deep learning, researchers resorted to neural networks methods (Ji and Eisenstein, 2015; Zhang et al., 2015; Chen et al., 2016; Liu et al., 2016b; Qin et al., 2016a; Liu and Li, 2016; Braud and Denis, 2016; Wu et al., 2016).

5.2 Multi-task learning

Multi-task learning framework adopts traditional machine learning with human-selected effective knowledge and the shared part is integrated into the cost function to prefer the main task learning. (Collobert and Weston, 2008) proposed a multi-task neural network trained jointly on the relevant tasks using weight-sharing (sharing the word embeddings with tasks). (Liu et al., 2016a) proposed the multi-task neural network by modifying the recurrent neural network for text classification tasks. (Lan et al., 2013) present a multi-task learning based system which can effectively use synthetic data for implicit discourse relation recognition. (Wu et al., 2016) use bilingually-constrained synthetic implicit data for implicit discourse relation recognition a multi-task neural network. (Liu et al., 2016b) propose a convolutional neural network embedded multi-task learning system to improve the performance of implicit discourse identification.

5.3 Deep learning with Attention

Recently deep learning with attention has been widely adopted by NLP researchers. (Zhou et al., 2016) proposed an attention-based Bi-LSTM for relation classification. (Wang et al., 2016c) proposed an attention-based LSTM for aspect-level sentiment classification. (Tan et al., 2016) proposed a attentive LSTMs for Question Answer Matching. (Wang et al., 2016a) proposed an inner attention based RNN (add attention information before RNN hidden representation) for Answer Selection in QA. (Wang et al., 2016b) proposed multi-level attention CNNs for relation classification. (Yin et al., 2016) proposed an attentive convolutional neural network for QA.

6 Concluding Remarks

We present a novel multi-task attention-based neural network model for implicit discourse relationship representation and identification. Our method captures both the discourse relationships through interactions between discourse arguments and the complementary knowledge through interactions between annotated and unannotated data. The experimental results showed that our proposed model outperforms the state-of-the-art systems on two benchmark corpora.

Acknowledgments

This work is supported by grants from Science and Technology Commission of Shanghai Municipality (14DZ2260800 and 15ZR1410700), Shanghai Collaborative Innovation Center of Trustworthy Software for Internet of Things (ZF1213) and NSFC (61402175).

References

Chloé Braud and Pascal Denis. 2015. Comparing word representations for implicit discourse relation classification. In *Empirical Methods in Natural Language Processing (EMNLP 2015)*.

Chloé Braud and Pascal Denis. 2016. Learning connective-based word representations for implicit discourse relation identification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 203–213, Austin, Texas. Association for Computational Linguistics.

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 ACL*.

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*, pages 160–167. ACM.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 12:2121–2159.

Robert Fisher and Reid Simmons. 2015. Spectral semi-supervised discourse relation classification. *Volume 2: Short Papers*.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, 3:329–344.

Man Lan, Yu Xu, and Zheng-Yu Niu. 2013. Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition. In *ACL (1)*, pages 476–485. Citeseer.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 343–351. Association for Computational Linguistics.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016a. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.

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*, pages 1224–1233, Austin, Texas. Association for Computational Linguistics.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016b. Implicit discourse relation classification via multi-task neural networks. *arXiv preprint arXiv:1603.02776*.

Emily Pitler, Annie Louis, and Ani 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: Volume 2-Volume 2*, pages 683–691. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The Penn Discourse TreeBank 2.0. In *LREC*. Citeseer.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016a. Implicit discourse relation recognition with contextaware character-enhanced embeddings. In *the 26th International Conference on Computational Linguistics (COLING), Osaka, Japan, December*.

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), Austin, USA, November*.

Attapol Rutherford and Nianwen Xue. 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proceedings of the NAACL-HLT*.

Attapol T Rutherford and Nianwen Xue. 2016. Robust non-explicit neural discourse parser in english and chinese. *ACL 2016*, page 55.

Ming Tan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2016. Improved representation learning for question answer matching. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 464–473, Berlin, Germany. Association for Computational Linguistics.

Bingning Wang, Kang Liu, and Jun Zhao. 2016a. Inner attention based recurrent neural networks for answer selection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1288–1297, Berlin, Germany. Association for Computational Linguistics.

Jianxiang Wang and Man Lan. 2016. Two end-to-end shallow discourse parsers for english and chinese in conll-2016 shared task. *Proceedings of the CoNLL-16 shared task*, pages 33–40.

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016b. Relation classification via multi-level attention cnns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1307, Berlin, Germany. Association for Computational Linguistics.

WenTing Wang, Jian Su, and Chew Lim Tan. 2010. Kernel based discourse relation recognition with temporal ordering information. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 710–719. Association for Computational Linguistics.

Yequan Wang, Minlie Huang, xiaoyan zhu, and Li Zhao. 2016c. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas. Association for Computational Linguistics.

Changxing Wu, xiaodong shi, Yidong Chen, Yanzhou Huang, and jinsong su. 2016. Bilingually-constrained synthetic data for implicit discourse relation recognition. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2306–2312, Austin, Texas. Association for Computational Linguistics.

Wenpeng Yin, Mo Yu, Bing Xiang, Bowen Zhou, and Hinrich Schütze. 2016. Simple question answering by attentive convolutional neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1746–1756, Osaka, Japan. The COLING 2016 Organizing Committee.

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*, pages 2230–2235.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. 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)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.

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*, pages 1507–1514, Stroudsburg, PA, USA. Association for Computational Linguistics.