# Towards Selective Data Enhanced Implicit Discourse Relation Recognition via Reinforcement Learning

Meilin Zhou*†, Qi Liang*†, Lu Ma*†, Dan Luo*†, Peng Zhang*†, Bin Wang‡
*Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
†School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
‡Xiaomi AI Lab, Beijing, China
{zhoumeilin,liangqi,malu,luodan,pengzhang}@iie.ac.cn, wangbin11@xiaomi.com

***Abstract*—As a fundamental task in NLP, recognizing implicit discourse relations remains a challenging problem for years. One of the most important reasons is the limited amounts of annotated data. On one hand, most existing methods use multi-task methods to enlarge data. External datasets are often fully introduced to the training process which may lead to a negative transfer of noisy data. On the other hand, some previous researches on selecting data mostly focus on designing rules with heuristic methods. It should be difficult to cover all aspects of good samples. Another drawback is that data selection can hardly generally fit well in any other recognition model since it was specialized with a specific discourse relation classifier.**

**In this paper, we propose a novel selective data enhanced model (SDE) based on reinforcement learning. Our model is a general framework, composed of two parts: 1) Discourse relation classifier is designed to identify relations, including multi-level representation module and relation recognition module. 2) Pseudo labeled data selector is designed to pick out data that can enhance the discourse relation classifier. We conduct joint learning alternately to optimize both of the classifier and the selector. Our model is able to expand data selectively. And classifier part can be replaced with any other complicated networks. To further exploit interact signals between arguments, we also present a multi-level representation based on BERT. Experiments show that our model achieves better performance than state-of-the-art methods.**

***Index Terms*—Discourse Relation Recognition, Data Enhance, Reinforcement Learning**

## I. Introduction

Implicit discourse relation recognition is intended to identify the relation between two adjacent arguments without explicit connectives such as *but*, *so*. It is an important step in semantic understanding and language generation, especially in many downstream NLP tasks, such as reading comprehension [2], topic segmentation [3], automatic abstract summarization [4] [5]. However, implicit discourse relation recognition remains a challenging task.

Existing works have made some attempts by expanding data to break through the limitations of small datasets, especially in neural-based methods. Considering the workload of manually labeling, previous studies expanded data from different perspectives. For example, Lan et al. (2013) [6] proposed a multi-task learning method to utilize synthetic data, which is automatically generated from unlabeled explicit data by dropping connectives. Synthetic data is regarded as pseudo labeled data. What existing similar works have in common is that all pseudo labeled data is feed to the whole training process. Previous researches [7] also have shown that there is a certain difference of linguistic dissimilarity between explicit and implicit data, i.e, pseudo labeled data cannot be equal with labeled data. Moreover, synthetic pseudo label data is naturally noisy. So, it results in that synthetic pseudo labeled data from explicit data cannot perform as well as natural implicit data. In fact, there have been some articles exploring the selection of good samples. Wang et al. (2012) [8] designed a heuristic single centroid clustering algorithm to select typical training samples. Rutherford et al. (2015) [9] selected samples by classifying explicit discourse connectives.

An obvious challenge of those works is that heuristic selecting methods cannot cover every aspect of high-quality data comprehensively. The performance of the selected data can only be manually tested. Also, no further updates can be made to the data selector even if we cannot select all good samples at one time. Since the data selector is specialized with a specific discourse relation classifier, most heuristic selecting methods cannot adjust adaptively in other discourse relation recognition methods. Whatś more, as we all know, the auto annotation model is trained with a small amount of existing news data. It is difficult to ensure high-quality samples can still be exactly picked out in heterogeneous datasets, such as chatting data. Therefore, itś necessary to perform an adaptive data selector to avoid negative effect accumulation caused by noise samples.

We provide two intuitive examples to illustrate why noisy samples can decline performance in implicit discourse relation recognition.

(1) Picked Example

Arg1 [Although dry beers still have limited distribution in the U.S. American brewers know it just a matter of time,]

Arg2 [(*before*) the Japanese begin exporting latest success in larger quantities.]

Pseudo label: *Comparison*

(2) Dropped Example

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Arg1 [After walking off the first flush of pain,]
Arg2 [(*however*) I finished the lesson.]
Pseudo label: *Temporal*

Here are two confusing examples with two possible senses *Comparison* and *Temporal*. After data selection, we picked out the first example with the correct pseudo label and drop the second one. The first one is possibly high quality because we got the correct pseudo label even though it has two connectives *although* and *before*. In contrast, the second example got the wrong pseudo label using the same data construction method. Itś difficult to distinguish the noticeable distinction between two examples using manually defined rules from heuristic methods.

In this paper, we proposed a general selective data enhance framework for implicit discourse relation recognition based on reinforcement learning. Our model framework consists of two modules: the discourse relation classifier and pseudo labeled data selector. Those two modules joint training iteratively. Our model can be adapted to any discourse relation classifiers, but not limited to a specific classifier. On one hand, policy function is the core component in pseudo labeled data selector. It enables the selector to select high-quality data. And selected samples are adopt to update parameters so as to get a better classification model. On the other hand, the performance of an updated classifier is fed to selector as a reward so as to guide policy function optimization by maximizing reward. Itś easy to assess the performance from the updated classifier base on the validation dataset. Through sufficient repetitively retraining, the data selection process and relation recognition process promote each other. This ultimately gets a better recognition model with higher quality samples picked out. The main contributions of our work are as follows:

- We propose a general reinforcement learning framework for implicit discourse relation recognition. This enables us to selectively expand data with high quality samples. Through data selection, it is able to prevent error accumulation of noise data. Experiments show that our method achieves better results than the state-of-the-art methods.
- We provide a multi-level representation based on BERT to catch arguments interacting information. Our work gets significant improvement by adding interact level representation.

## II. Related Work

### A. Discourse Relation Recognition

Since the release of PDTB 2.0 corpus [10], which is a benchmark corpus for discourse relations, there have been many studies recognizing discourse relation, mainly focusing on the more challenging task of implicit discourse relation classification [11]without explicit discourse connective information provided. Early studies [12] [13] [14] used a feature engineering approach, extracting traditional linguistic features from two discourse arguments such as polarity labels and lexical characteristics until Joonsuk al. (2012) [15] summarized and optimized the characteristics presented in previous studies. Among those, some studies also tried to introduce the interactive information between two arguments.

In recent years, deep learning methods have got significant progress in many fields such as machine translation and so on. Researches on implicit discourse relation recognition using neural-based methods also achieved good results. Some works are devoted to embedding improvement [16] in order to gain complicated text representations from different aspects. Qin et al. (2016) [17] optimized word-level representation exploiting context-aware character information. Bai and Zhao(2018) [18] gained deep representation from character, subword, word, sentence, and sentence pair levels. Other studies designed more complicated and improved neural network models, such as convolutional neural network (CNN) [19], multi-task method [20], attention mechanisms [21] and adversarial learning [22].

Neural-based methods contain a mass of model parameters compared with traditional feature-based machine learning methods. Thus, a common characteristic of neural-based methods are the high complexity of algorithms. The size of implicit instances in PDTB 2.0 is only 16,253 within a totally 40,600 annotated instances due to the manual annotation complexity. Given the limited amount of annotated data in comparison to the number needed, to avoiding data sparsity problem and taking advantage of the deep learning models, one potential method is to provide sufficient training data.

Previous researches have made some attempts to enlarge datasets exploiting both labeled and unlabeled data. And most of the existing works use multi-task methods. Hugo et al. [23] extended feature vectors with non-label data using co-occurrence information. Lan et al.(2013) [6] proposed a multi-task method combining implicit and explicit data. Another work [20] performed two representation learning by sharing parameters in two datasets. Liu et al.(2016) [24] designed a CNN embedded multi-task method to share representations in different tasks.

What they have in common is that all external data is feed to the whole training process and some external data is synthesized from explicit data. Previous studies [7] have shown that there is a certain difference in linguistic dissimilarity between explicit and implicit data. Thus, synthetic external data from explicit data cannot perform as well as annotated implicit data. It indicates that instead of simply adding all pseudo labeled data into model training, we should partly select high quality data into model training.

There have been some articles exploring the selection of good samples. Wang et al. (2012) [8] designed a heuristic single centroid clustering algorithm to select typical training samples. Rutherfordet al. (2015) [9] selected samples by classifying explicit discourse connectives according to omission rate and context differential of connectives. Existing works use heuristic approaches to select data that cannot cover every specific aspect comprehensively. Thus, we provide a reinforcement learning method to select good samples. Those selected samples have direct feedback on our final task, implicit relation recognition. The selection process and relation recognition promote each other by updating model parameters and sharing

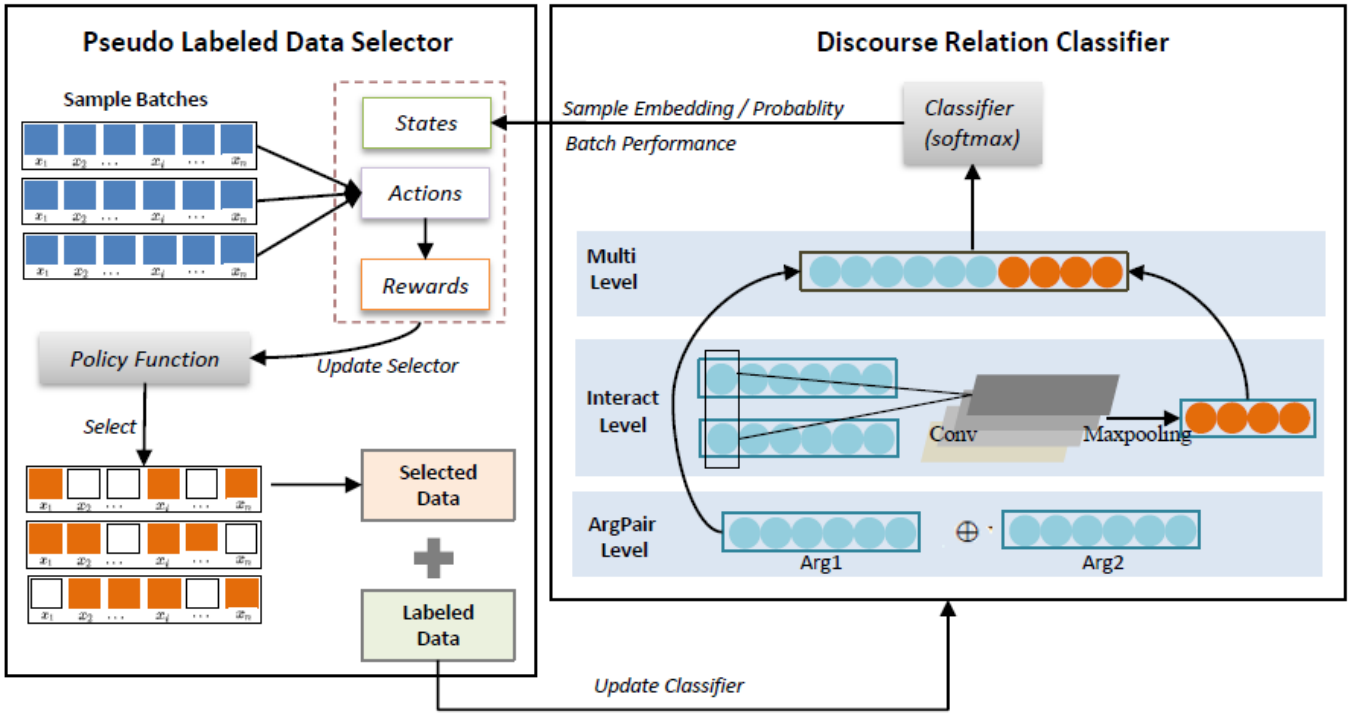


Fig. 1. Overall Methodology

representation during the joint training process iteratively. This ultimately gets better recognition model with higher quality samples picked out. Our model is a universal framework, allow us to select samples adaptively under different classifiers.

### B. Reinforcement Learning

Sutton et al. first proposed the concept of Reinforcement Learning in the late 1970s [25]. Depending on the specific problem and algorithm, Reinforcement Learning can be divided into several branches, such as Transfer Reinforcement Learning, Deep Reinforcement Learning, Multi-agent Reinforcement Learning and so on. With the success of DeepMind and AlphaGo [26], Reinforcement Learning [27] is gaining more attention and has been used in tasks including machine reading comprehension [28], entity relationship classification [29] and text matching [30].

From an optimization point of view, Reinforcement Learning can be summarized as value-based approaches and policy-based approaches. Our work focus on designing policy-based approaches to select external pseudo label data. According to our policy function, we can decide whether to choose one sample at the current state. Finally, we get a maximum reward from all selected samples.

## III. Model

### A. Problem Definition

For a given set of implicit samples $X = \{x_1, \ldots, x_n\}$. Each sample $x$ is composed of two arguments $Arg1$, $Arg2$, implicit discourse relation recognition is to predict the relation $r$ between $Arg1$ and $Arg2$.

### B. Methodology Overview

As described in Fig.1, our model consists of two components: discourse relation classifier and pseudo labeled data selector. As mentioned earlier, our model can be integrated into any discourse relation classifiers. In this paper, we use a basic CNN architecture to recognize discourse relations. In pseudo labeled data selector, data selection is essentially a serialized decision making process for each batch of data. At each state, data selector will give out the corresponding action which reflects whether to select a sample or not. Once a batch of data finished getting the corresponding actions, they are provided to the discourse relation classifier to obtain a delayed reward on validation data. Then, the reinforced selector will be updated based on the policy gradient method, and the discourse relation classifier will be trained based on labeled and selected data. The whole training process is shown in Algorithm 1.

### C. Discourse Relation Classifier

We adopt the basic CNN architecture to recognize discourse relations, including multi-level representation and relation recognition. Our representations are rendered in multi-levels, including an argument pair level and interact level representation. Argument pair level representation is obtained from pre-trained BERT base model. Interact level representation is calculated through convolution on argument pair level. By concating argument pair level representation and interactive representation, we can get multi-level representation. Our model is a general framework, key work is focused on the

**Algorithm 1** Overall Training Process

1: Data Construction: Automatic tagging on unlabeled dataset, filtering samples with connectives and extract pseudo labeled samples;
2: Initial Embedding: Get initial BERT base embeddings of each sample in both labeled implicit dataset $E_{imp}$, labeled explicit dataset $E_{exp}$ and pseudo labeled dataset $E_{pseu}$ distributed in several batches;
3: Pretrain discourse relation classifier to get original classifier model $C_0$, using embedded labeled data $E_{imp}$;
4: Pretrain policy function of data selector to get original policy function $\pi_0$, using multi-level embedding and caculating rewards of candidates according to $C_0$;
5: Joint training of discourse relation classifier and data selector.

selection part, so we use the basic CNN model in discourse relation classifier as a representative method.

BERT got great success in many tasks [31] [32] since it was proposed [33] by Google in 2018. It is a bidirectional encoder representation of Transformers. Many recent NLP architectures, training methods, and language models are trying to propose optimized models , such as SpanBERT, RoBERTa [34]. Inspired by those research results, we adopted BERT pre-trained models to get original embeddings. Each sample contains two arguments and each argument is embedded to 768 dimensions. Considering that we can get interaction information between the two arguments $Arg_1$ and $Arg_2$ through the convolution operation, because convolution brings a multiplication between input matrix and convolution kernel. At last, a combined multi-level representation is fed into CNN network.

Given an original BERT representation $(E_x^1, E_x^2)$ of one sample $x$, We feed these original representations into CNN convolution layer and maxpooling layer to get a convolution representation $O_x = \max\_\text{pool}(\text{conv}(x))$. Through convolution operation, dot multiplying between convolution kernel and one specific dimension makes it possible for two arguments to interact with each other. Then, we get a combined representation $R_x = O_x \oplus E_x^1 \oplus E_x^2$ by concating the original argument pair level representation with the convolutional interact level representation. Finally, the combined multi-level representation is given to a non-linear layer and a fully connected layer.

Our discourse relation classifier is based on the combined multi-level representation. The probability of relation is defined as follows:

$$p(r|x;\mathbf{\Phi}) = \text{softmax}\left(\mathbf{W}_r * \tanh(E_x) + \mathbf{b}_r\right) \tag{1}$$

where $\mathbf{W}_r$ and $\mathbf{b}_r$ are weight matrix and bias vector in fully connected layer, $r$ refers to the relation label, $x$ is the sample. And $\mathbf{\Phi}$ refers to the collection of classifier parameters.

**Optimization.** For a given set of training samples $X = \{x_1, x_2..., x_n\}$, we aim to get maximum likelihood. We adopt multi-class cross entropy as the loss function as follows:

$$\mathcal{J}(\Phi) = -\frac{1}{|X|}\sum_{i=1}^{|X|}\log p\left(r_i|x_i;\Phi\right) \tag{2}$$

To get a minimized loss function, AdaGrad is used to optimize this objective function.

### *D. Pseudo Labeled Data Selector*

Pseudo Labeled Data Selector is implemented based on a reinforcement learning framework. For each batch, data selection is a sequential decision making problem. Action, state, and reward are three major elements. We will introduce them separately in this section.

**Action.** We define action as $a_i \in \{0, 1\}$, 1 means to select and 0 means to drop the sample from a given state in one batch, which is regarded as the decision. In each state of samples sequence, $a_i$ is calculated according to a learned policy network function $\pi_\Theta$. We define $\pi_\Theta$ as follows:

$$\pi_\Theta(s_i) = P_\Theta\left(a_i|s_i\right) = \text{softmax}\left(\mathrm{W} * s_i + \mathrm{b}\right) \tag{3}$$

where $P_\Theta$ gives out the probability distribution, $\mathbf{W}$ and $\mathbf{b}$ are weight matrix and bias vector in fully connected layer, $s_i$ refers to the state vector of the i-th sample in one sequence, $\Theta$ is the collection of parameters.

**State.** We designed a state function, which contains representation of current sample, likelihood probability on pseudo label and onehot vector of the sampleś pseudo label. When making decision on the i-th sample in one batch, state is denoted as $s_i$.

$$s_i = E_i \oplus p(r_i|x_i) \oplus l_i \tag{4}$$

where $e_i$ is the multi-level representation of sample $x_i$, this is designed to capture the consistency between current sample and implicit samples in the expected category. $p(r_i|x_i)$ refers to probability on pseudo label which is calculated from the discourse relation classifier. $l_i$ is the onehot vector of the pseudo label of sample $x_i$. So we can get the pseudo label and its corresponding probability. If we get high probability in pseudo label, it is quite possible that the pseudo label is correct. Therefore, probability reflects the quality of the pseudo labels. In this way, state function allows us to select those samples with similar representations to the real implicit annotated data. Moreover, selected samples are labeled with correct pseudo category as closely as possible.

**Reward.** We sample the actions of each state in one batch sequentially according to the policy function. The batch collection $B_b$ is composed by $\left\{x_1, \ldots, x_{|B_b|}\right\}$. Considering that reward function shows whether this chosen sample will be good for discourse relation classifier. We update the discourse relation classifier with $\hat{B}_b$, which refers to the selected samples for a given batch $B_b$. Then we obtain the performance of the updated discourse relation classifier on validation dataset as reward of batch $B_b$, formulated with $R_b$ .

However, $R_b$ is the preliminary batch reward. We integrate discounted rewards of other batches as the total future reward. We expect that earlier selected samples work better, as data

selection is a serialized decision making process on each batch. Therefore, weight of rewards in different batches declined gradually. The total future reward of batch $B_b$ is designed as follows:

$$Q_b = \sum_{t=0}^{N-b} \gamma^t R_{b+t} \tag{5}$$

where $N$ is the total batch number, $\gamma$ is discount factor of rewards.

**Optimization.** The ultimate goal of policy network is to maximize expected total reward in each episode with several batches. Our objective function is defined as follows:

$$\mathcal{J}(\Theta) = E_{\pi_\ominus}\left[\sum_{b=1}^{N} R_b\right] \tag{6}$$

where $R_b$ is the batch reward , $N$ is the number of batches in each episode. We use the policy gradient to optimize the objective function. Parameters $\Theta$ was updated according to gradient like:

$$\Theta \leftarrow \Theta + \alpha \sum_{i=1}^{|B_b|} v_i \nabla_\Theta \log \pi_\Theta\left(\mathrm{s}_i\right) \tag{7}$$

where $\alpha$ is the learning rate, $s_i$ refers to the state of the i-th action in a batch $B_b$. At any states $s_i$ of a given batch, the action reward equals to the batch reward, i.e. $v_i = Q_b$.

### *E. Joint Learning*

Considering that discourse relation classifier and pseudo labeled data selector interact with each other closely, we conduct a joint learning during the whole training process to optimize models. As mentioned, discourse relation classifier and data selector should be pretrained separately before we start joint training. Firstly, we pretrain the discourse relation classifier to ensure that rewards in data selector can be acquired. Then, we pretrain the policy function in data selector in order to select training data. The entire joint training process can be described like Algorithm2.

TABLE I
THE DISTRIBUTION OF DISCOURSE RELATION IN PDTB AND BLLIP

| | **PDTB(imp)** | **PDTB(exp)** | **BLLIP** |
|---|---|---|---|
| Comparison | 2231 | 5471 | 62484 |
| Expansion | 3844 | 3250 | 65781 |
| Contingency | 7999 | 6298 | 66674 |
| Temporal | 787 | 3440 | 57524 |
| Total | 14861 | 18459 | 252463 |

## IV. DATASET CONSTRUCTION

In this paper, we use two corpora: PDTB 2.0 [10], known as the largest annotated discourse corpus containing 40,600 instances from Wall Street Journal articles. BLLIP corpus [35] is unlabeled North American News Text and we use an automatic annotation method to synthetic pseudo labeled discourse relations.

**Algorithm 2** Joint Learning between Discourse Relation Classifier and Pseudo Labeled Data Selector

**Input:**

Episode L; The embeddings of labeled implicit dataset $E_{imp}$, composed of training/dev sets; The embeddings of pseudo labeled dataset $E_{pseu}$, all of those are splited into batches $B = (B_1, \ldots, B_N)$, each batch $B_b$ is composed by $\{x_1, \ldots, x_{|B_b|}\}$;

The pretrain discourse relation classifier $C_0$;

The pretrained data selector policy function $\pi_0$ according to $C_0$;

**for** each episode $l \in L$ **do**

  **for** each batch $B_b$ **do**

    1.Caculate states using Equation (4) according to new embeddings acquired from latest classifier;

    2.Obtain sampled action sequence for every state in one batch according to selector policy function $\pi_{l-1}$, so we get several $\langle s_i, a_i \rangle$ pairs for training;

    3. Get selected data collection $\hat{B}_b$ according to actions.

    4. Evaluate the performance on validation dataset using discourse relation classifier which is trained with $\hat{B}_b$ , to get corresponding rewards $Q_b$ as in Equation (5);

  **end for**

  1.Update params $\Theta$ in data selector agent policy function $\pi_l$ following Equation (3) with training set $\{\langle s_i, a_i \rangle\}$ of all batches;

  2.Select samples of $B$ using updated data selector $\pi_l$;

  3.Use new selected training data and PDTB implicit data to train a new get $C_l$, update params $\Phi$;

  4.Produce new embedding of each $x_i \in B$;

**end for**

Previous work [9] has proven that pseudo labeled data can provide strong additional signal to implicit relation classifier, despite that contains a small number of false labels. Rutherford [9] use heuristics method to define omissible rate on explicit discourse connectives to collect discourse relations. Another work in Open-domain Dialogues [36] extract relations use connective words that only appear in one class according to statistical analysis.

We therefore construct pseudo labeled BLLIP. We generate it by automatically annotating unlabeled data. Based on this pseudo labeled data, we can then extract the explicit discourse relation argument pairs. The original BLLIP contains over 1.3 million sentences. Automatic labeling is divided into three steps. Firstly, we filter out sentences that contain any selected connectives follow their work [36]. The selected connectives include: *but, however, although, by contrast, because, so, thus, as a result, consequently, therefore, also, for example, in addition, instead, indeed, moreover, for instance, in fact, furthermore, or, then, previously, earlier, later, after, before*. Secondly, we train an explicit classifier on PDTB explicit data and use it to label the selected sentences with four level-1 discourse relations(Comparison, Expansion, Contingency and Temporal). Finally, we extract sample sentences

TABLE II
4-WAY CLASSIFICATION RESULTS IN DIFFERENT DATA SETS USING PRECISION, RECALL AND F1

| | | **Baseline** | **None Selected Data Enhance** | | **Selected Data Enhance** | |
|---|---|---|---|---|---|---|
| | | ***PDTB*** | ***PDTB+PDTB(exp)*** | ***PDTB+BLLIP*** | ***PDTB+PDTB(exp)*** | ***PDTB+BLLIP*** |
| Comp. | P | 42.31 | 27.69 | 25.55 | 40.82 | **44.85** |
| | R | 37.67 | 58.22 | 23.97 | 41.10 | **41.78** |
| | F1 | 39.86 | 37.53 | 24.73 | 40.96 | **43.26** |
| Cont. | P | 44.41 | **51.92** | 38.98 | 45.78 | 46.78 |
| | R | 47.46 | 29.35 | 58.33 | 60.87 | **68.48** |
| | F1 | 45.88 | 37.5 | 46.73 | 52.26 | **55.59** |
| Expa. | P | 66.26 | 69.61 | 63.07 | 71.40 | **74.50** |
| | R | **68.17** | 55.22 | 42.09 | 60.61 | 59.89 |
| | F1 | **67.20** | 61.59 | 50.49 | 65.56 | 66.40 |
| Temp. | P | 36.73 | 24.65 | 19.20 | 38.33 | **42.37** |
| | R | 26.47 | 51.47 | 35.29 | 33.82 | **36.76** |
| | F1 | 30.77 | 33.34 | 24.87 | 35.93 | **39.37** |
| Macro-F | | 45.92 | 42.49 | 36.71 | 48.68 | **51.16** |

into (Arg1,Connecitve, Arg2). After dropping the discourse connectives, we should be able to treat them as additional implicit discourse instances and we get totally 0.25 million pseudo labeled samples.

Distribution of the annotated discourse relation is shown in Table 1. It is easy to see that the number of new dataset is 10 times higher than PDTB, which should be more useful for neural-based algorithms.

## V. EXPERIMENTS

### A. Experimental Setting

**Data Setting.** We follow the standard settings for the PDTB v2.0 dataset(Sections 2-20, Sections 0-1 and Sections 21-22 for training, development and testing), known as PDTB-Ji. To be clear, PDTB in experiment table refers to the implicit dataset and PDTB(exp) is the explicit dataset with connectives dropped. BLLIP is the external pseudo labeled dataset with connectives dropped as described in Dataset Construction section.

**Representation.** We adopt BERT pre-trained models to get original sentence embeddings of two arguments. Each argument is embedded to 768 dimensions. In order to get interaction information of two arguments, we set convolution kernel size with [2,200]. After max-pooling, we get an interaction value from each convolution filter. In this paper, we give 290 filters in convolutional layer.

**Training.** Discourse relation classifier and pseudo labeled data selector should be pretrained separately before we start joint training. Original classifier is trained with PDTB implicit data and original selector is trained with target selecting data. Then, we update the parameters of classifier and policy function iteratively. For classifier training process, we use a multiclass cross-entropy loss, optimized with AdaGrad follow previous studies. For selector training process, we aim to get a maximum expect total reward, also optimized with AdaGrad algorithm.

**Parameters.** In discourse relation classifier, we set cnn kernal-size as [2, 200] and use 290 convolution filters. Thus, the sample embedding dimension is 1826. We set dropout as 0.9 and learning rate is 0.001. In pseudo labeled data selector, the training episode is fixed to 30 and learning rate is set as 0.002. We employ reward function with a discount of 0.8.

**Evaluation.** To evaluate the implicit discourse relation recognition performance, we adopt F1 for one-vs.-others binary classification. Macro-averaged F1 is used to evaluate 4-way classification and Precision, Recall, F1 are utilized for assessing each class. To be clear, we use P, R, F1 in tables.

### B. Selective Data Enhance Performance

To evaluate the performance of reinforcement learning framework, namely, to verify the quality of the selected data, we design a group of experiments. We conduct experiments on PDTB annotated data as a baseline. To be fair, classifier is same with that in SDE model. The four-way classification results are showed in Table 2. We find that with fully PDTB explicit data expanded, we get a lower F1 compared with PDTB implicit data only. This might be caused by the certain difference of linguistic dissimilarity between explicit and implicit data. On the contrary, the four-way classification results show a significant decrease when adding all external BLLIP data without data selection. The decline in performance is mainly due to noises in automatic labels within large amount of data. However, with selected samples, we get an obvious outstanding result. This should be an effective signal of our SDE model which means we can get certain high-quality data for implicit discourse relation recognition through selecting.

### C. Comparison with the State-of-the-art

In this section, we compare our reinforced SDE model with state-of-the-art baselines. Let's have a brief overview to baseline models. In Table 4, we categorize baselines as three types, selective methods, parsing networks and representation methods.

Selective methods mainly focus on selecting good samples for training process in discourse classifier. Wang et al. (2012) [8] used clustering algorithm to select typical samples and Rutherford and Xue (2015) [9] utilized discourse connectives properties to gather extra weakly labeled data. Our model gets obviously higher performance than those selective methods. This is mainly due to the reward function's ability to directly

TABLE III
PERFORMANCES OF DIFFERENT SYSTEM USING F1

| | Binary | | | | Four-way |
|---|---|---|---|---|---|
| | ***Comp.*** | ***Cont.*** | ***Expa.*** | ***Temp.*** | |
| Wang et al. (2012) | 28.5 | 48.5 | 71.1 | 14.7 | 40.2 |
| Rutherford and Xue (2015) | 41.0 | 53.8 | 69.4 | 33.3 | 40.5 |
| Lan et al. (2013) | - | - | - | - | 44.64 |
| Liu and Li (2016a) | 37.91 | 55.88 | 69.97 | 37.17 | 44.98 |
| Liu and Li (2016b) | 39.86 | 53.69 | 69.71 | 37.61 | 44.95 |
| Qin et al. (2017) | 40.87 | 54.56 | 72.38 | 36.20 | - |
| Lan et al. (2017) | 40.73 | **58.96** | **72.47** | 38.50 | 47.80 |
| Lei et al. (2017) | 40.47 | 55.36 | 69.50 | 35.34 | 46.46 |
| Lei et al. (2018) | 43.24 | 57.82 | 72.88 | 29.10 | 47.15 |
| Bai and Zhao (2018) | **47.85** | 54.47 | 70.60 | 36.87 | 51.06 |
| SDE | 40.0 | 55.04 | 71.45 | **39.52** | **51.16** |

reflect contributions of selected samples in discourse relation classifier. Selecting data in this way does not require consideration of the characteristics of high-quality data in all aspects.

Another genres mostly work on designing complicated parsing networks, including multi-task methods, attention network and adversarial models. Part of those researches is aiming to introduce external data. Lan et al. (2013) [6] designed an auxiliary task utilizing explicit PDTB data and synthetic data. Liu and Li (2016a) [24] used multi-task method to synthesize tasks by focus on learning both unique and shared representations. Liu and Li (2016b) [37] designed multi-level attention networks (NNMA). Qin et al. (2017) [22] is an adversarial model to enable competition between the implicit network and a rival feature discriminator. Lan et al. (2017) [20] used a multi-task method based on annotated and unlabeled data. Our model gets higher results than those methods, especially compared with those using external data. To be mentioned,our model is a general framewok, classifier part can be replaced with any other better-performing networks in the future to get higher performance.

Representation methods generally combined with utilizing rich representation features. Lei et al.(2017) [38] demonstrated embeddings that considering word semantic interaction. Lei et al. (2018) used linguistic properties in classification as complex features. Bai and Zhao (2018) [18] designed a representation model with different text levels. Compared with representation methods, we achieve better performance than SWIM and linguistic properties mattered recognition methods. Experiment results also reflect the effectiveness of interact level representation.

In conclusion, compared with state-of-the-art, our model gets obviously better results than selective methods and parsing networks. And we verified the sematic interaction between two arguments.

### *D. Embedding Performance on Different Representation Levels*

Besides, we conduct experiments to verify the effectiveness of representations in different levels. We implement a baseline representation (Word2Vec) which directly obtain the output vectors of two arguments. Then we consider using pre-trained BERT base model to get representations in argument pair level. To further exploit interaction information between two arguments, an interact representation is obtained by convolution. We also introduced a multi-level representation through concating BERT base representation and the interactive representation on experiments. To directly reflect the representations influence, we adopt a simple perceptron architecture in which different embeddings represent sentence arguments. The detailed results are shown in Table 4. We conduct four-way classification on implicit discourse relations in PDTB-Ji. Itś obvious that BERT did great job in representation compared with Word2Vec. This confirms BERT's contribution to NLP tasks, especially for the representation of sentences from a semantic level. Standing on the shoulders of giants, semantic interaction can be naturally catches in our interactive level representation. By adding interactive level representation, we get better results, which proved the semantic relevance between two arguments.

TABLE IV
EMBEDDING PERFORMANCE ON THE 4-WAY CLASSIFICATION USING PRECISION, RECALL AND F1

| | | Baseline | BERT-base Representation | | |
|---|---|---|---|---|---|
| | | ***Word2Vec*** | ***ArgPair Level*** | ***Interact Level*** | ***Multi Level*** |
| Comp. | P | **54.30** | 41.67 | 36.43 | 42.31 |
| | R | **71.58** | 30.82 | 32.19 | 37.67 |
| | F1 | **61.75** | 35.43 | 34.18 | 39.86 |
| Cont. | P | 16.49 | 43.93 | **45.78** | 44.41 |
| | R | 10.96 | 38.04 | 41.30 | **47.46** |
| | F1 | 13.17 | 40.77 | 43.42 | **45.88** |
| Expa. | P | 35.94 | 63.06 | 65.19 | **66.26** |
| | R | 8.33 | **75.54** | 74.10 | 68.17 |
| | F1 | 13.53 | 68.74 | **69.36** | 67.20 |
| Temp. | P | 9.87 | 36.36 | 30.56 | **36.73** |
| | R | 22.06 | 17.65 | 16.18 | **26.47** |
| | F1 | 13.64 | 23.76 | 21.16 | **30.77** |
| Macro-F | | 25.52 | 42.18 | 42.03 | **45.92** |

## VI. CONCLUSION AND FUTURE WORK

We present a novel general framework based on reinforcement learning to select high-quality samples for implicit discourse relation classification. In our model, selected samples are used to retrain the classifier. According to the updated classifier, data selector which is formally an agent in reinforcement learning can be repeat updated again. The advantage of our approach lies in the ability to examine the performance feedback of selected data and the selector framework is adaptive in different classifiers. Experiment results show our proposed model is able to achieve state-of-the-art F1 scores, accompanied by generating some high quality pseudo label data. Our work proved that pre-trained BERT has a significant effect and we find evidences that two arguments have sematic relevance. In the future we plan to explore a semi-supervised data generation method to reduce noise ratio in pseudo labeled data.

## ACKNOWLEDGMENT

This work is supported by the National Key R&D Program with No.2016QY03D0503, 2016YFB081304, Strategic Priority Research Program of Chinese Academy of Sciences, Grant No.XDC02040400, National Natural Science Foundation of China (No.61602474, No.61602467, No.61702552). Q. Liang is the corresponding author.

## REFERENCES

[1] J. G. Barnitz, "Toward understanding the effects of cross-cultural schemata and discourse structure on second language reading comprehension," *Journal of reading behavior*, vol. 18, no. 2, pp. 95–116, 1986.

[2] J. Clarke and M. Lapata, "Modelling compression with discourse constraints," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 1–11.

[3] P. Cardoso, M. Taboada, and T. Pardo, "On the contribution of discourse structure to topic segmentation," in *Proceedings of the SIGDIAL 2013 Conference*, 2013, pp. 92–96.

[4] S. Gerani, Y. Mehdad, G. Carenini, R. T. Ng, and B. Nejat, "Abstractive summarization of product reviews using discourse structure," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1602–1613.

[5] A. Cohan, F. Dernoncourt, D. S. Kim, T. Bui, S. Kim, W. Chang, and N. Goharian, "A discourse-aware attention model for abstractive summarization of long documents," *arXiv preprint arXiv:1804.05685*, 2018.

[6] M. Lan, Y. Xu, and Z. Niu, "Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2013, pp. 476–485.

[7] C. Sporleder and A. Lascarides, "Using automatically labelled examples to classify rhetorical relations: An assessment," *Natural Language Engineering*, vol. 14, no. 3, pp. 369–416, 2008.

[8] X. Wang, S. Li, J. Li, and W. Li, "Implicit discourse relation recognition by selecting typical training examples," in *Proceedings of COLING 2012*, 2012, pp. 2757–2772.

[9] 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.

[10] 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*. Citeseer, 2008.

[11] 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.

[12] Z. Lin, M.-Y. Kan, and H. T. Ng, "Recognizing implicit discourse relations in the penn discourse treebank," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2009, pp. 343–351.

[13] A. Louis, A. Joshi, R. Prasad, and A. Nenkova, "Using entity features to classify implicit discourse relations," in *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 2010, pp. 59–62.

[14] W. Wang, J. Su, and C. L. Tan, "Kernel based discourse relation recognition with temporal ordering information," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 710–719.

[15] J. Park and C. Cardie, "Improving implicit discourse relation recognition through feature set optimization," in *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 2012, pp. 108–112.

[16] 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.

[17] L. Qin, Z. Zhang, and H. Zhao, "Implicit discourse relation recognition with context-aware character-enhanced embeddings," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 1914–1924.

[18] H. Bai and H. Zhao, "Deep enhanced representation for implicit discourse relation recognition," *arXiv preprint arXiv:1807.05154*, 2018.

[19] B. Zhang, J. Su, D. Xiong, Y. Lu, H. Duan, and J. Yao, "Shallow convolutional neural network for implicit discourse relation recognition," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2230–2235.

[20] 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.

[21] A. Cianflone and L. Kosseim, "Attention for implicit discourse relation recognition," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[22] L. Qin, Z. Zhang, H. Zhao, Z. Hu, and E. P. Xing, "Adversarial connective-exploiting networks for implicit discourse relation classification," *arXiv preprint arXiv:1704.00217*, 2017.

[23] H. Hernault, D. Bollegala, and M. Ishizuka, "A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 399–409.

[24] 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.

[25] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[26] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge," *nature*, vol. 550, no. 7676, pp. 354–359, 2017.

[27] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[28] S. Wang, M. Yu, X. Guo, Z. Wang, T. Klinger, W. Zhang, S. Chang, G. Tesauro, B. Zhou, and J. Jiang, "R 3: Reinforced ranker-reader for open-domain question answering," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[29] J. Feng, M. Huang, L. Zhao, Y. Yang, and X. Zhu, "Reinforcement learning for relation classification from noisy data," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[30] C. Qu, F. Ji, M. Qiu, L. Yang, Z. Min, H. Chen, J. Huang, and W. B. Croft, "Learning to selectively transfer: Reinforced transfer learning for deep text matching," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 2019, pp. 699–707.

[31] C. Sun, L. Huang, and X. Qiu, "Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence," *arXiv preprint arXiv:1903.09588*, 2019.

[32] X. Li, Z. Zhang, W. Zhu, Z. Li, Y. Ni, P. Gao, J. Yan, and G. Xie, "Pingan smart health and sjtu at coin-shared task: utilizing pre-trained language models and common-sense knowledge in machine reading tasks," in *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, 2019, pp. 93–98.

[33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[34] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[35] D. McClosky, E. Charniak, and M. Johnson, "Bllip north american news text," 2008.

[36] M. D. Ma, K. K. Bowden, J. Wu, W. Cui, and M. Walker, "Implicit discourse relation identification for open-domain dialogues," *arXiv preprint arXiv:1907.03975*, 2019.

[37] Y. Liu and S. Li, "Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention," *arXiv preprint arXiv:1609.06380*, 2016.

[38] W. Lei, X. Wang, M. Liu, I. Ilievski, X. He, and M.-Y. Kan, "Swim: A simple word interaction model for implicit discourse relation recognition."