Adversarial Connective-exploiting Networks for Implicit Discourse Relation Classification

Lianhui Qin[1,2], **Zhisong Zhang**[1,2], **Hai Zhao**[1,2,*], **Zhiting Hu**[3], **Eric P. Xing**[3]

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
[3]Carnegie Mellon University

{qinlianhui, zzs2011}@sjtu.edu.cn,   zhaohai@cs.sjtu.edu.cn,
{zhitingh, epxing}@cs.cmu.edu

Abstract

Implicit discourse relation classification is of great challenge due to the lack of connectives as strong linguistic cues, which motivates the use of annotated implicit connectives to improve the recognition. We propose a feature imitation framework in which an implicit relation network is driven to learn from another neural network with access to connectives, and thus encouraged to extract similarly salient features for accurate classification. We develop an adversarial model to enable an adaptive imitation scheme through competition between the implicit network and a rival feature discriminator. Our method effectively transfers discriminability of connectives to the implicit features, and achieves state-of-the-art performance on the PDTB benchmark.

1 Introduction

Discourse relations connect linguistic units such as clauses and sentences to form coherent semantics. Identification of discourse relations can benefit a variety of downstream applications including question answering (Liakata et al., 2013), machine translation (Li et al., 2014), text summarization (Gerani et al., 2014), opinion spam detection (Chen and Zhao, 2015), and so forth.

Connectives (e.g., *but*, *so*, etc) are one of the most critical linguistic cues for identifying discourse relations. When explicit connectives are present in the text, a simple frequency-based mapping is sufficient to achieve over 85% classification accuracy (Xue et al., 2016; Li et al., 2016). In contrast, *implicit discourse relation recognition* has long been seen as a challenging problem, with the best accuracy so far still lower than 50% (Chen et al., 2015). In the implicit case, discourse relations are not lexicalized by connectives, but to be inferred from relevant sentences (i.e., *arguments*). For example, the following two adjacent sentences Arg1 and Arg2 imply relation *Cause* (i.e., Arg2 is the cause of Arg1).

> [*Arg1*]: Never mind.
>
> [*Arg2*]: You already know the answer.
>
> [*Implicit connective*]: Because
>
> [*Discourse relation*]: Cause

Various attempts have been made to directly infer underlying relations by modeling the semantics of the arguments, ranging from feature-based methods (Lin et al., 2009; Pitler et al., 2009) to the very recent end-to-end neural models (Chen et al., 2016a; Qin et al., 2016c). Despite impressive performance, the absence of strong explicit connective cues has made the inference extremely hard and hindered further improvement. In fact, even the human annotators would make use of connectives to aid relation annotation. For instance, the popular Penn Discourse Treebank (PDTB) benchmark data (Prasad et al., 2008) was annotated by first inserting a connective expression (i.e., *implicit connective*, as shown in the above example) manually, and determining the abstract relation by combining both the implicit connective and contextual semantics.

Therefore, the huge performance gap between explicit and implicit parsing (namely, 85% vs 50%), as well as the human annotation practice, strongly motivates to incorporate connective information to guide the reasoning process. This paper aims to advance implicit parsing by making use of annotated implicit connectives available in training data. Few recent work has explored such combination. Zhou et al. (2010) developed a two-step approach by first predicting implicit connectives whose sense is then disambiguated to obtain the relation. However, the pipeline approach usually suffers from error propagation, and the method itself has relied on hand-crafted features which do not necessarily generalize well. Other research leveraged explicit connective examples for data augmentation (Rutherford and Xue, 2015; Braud and Denis, 2015; Ji et al., 2015; Braud and Denis, 2016). Our work is orthogonal and complementary to this line.

In this paper, we propose a novel neural method that incorporates implicit connectives in a principled *adversarial* framework. We use deep neural models for relation classification, and take the intuition that, sentence arguments integrated with connectives would enable highly discriminative neural features for accurate relation inference, and an ideal implicit relation classifier, even though without access to connectives, should mimic the connective-augmented reasoning behavior by extracting similarly salient features. We therefore setup a secondary network in addition to the implicit relation classifier, building upon connective-augmented inputs and serving as a feature learning model for the implicit classifier to emulate.

Methodologically, however, feature imitation in our problem is challenging due to the semantic gap induced by adding the connective cues. It is necessary to develop an adaptive scheme to flexibly drive learning and transfer discriminability. We devise a novel adversarial approach which enables a self-calibrated imitation mechanism. Specifically, we build a discriminator which distinguishes between the features by the two counterpart networks. The implicit relation network is then trained to correctly classify relations and simultaneously to fool the discriminator, resulting in an adversarial framework. The adversarial mechanism has been an emerging method in different context, especially for image generation (Goodfellow et al., 2014) and domain adaptation (Ganin

et al., 2016; Chen et al., 2016c). Our adversarial framework is unique to address neural feature emulation between two models. Besides, to the best of our knowledge, this is the first adversarial approach in the context of discourse parsing. Compared to previous connective exploiting work (Zhou et al., 2010; Xu et al., 2012), our method provides a new integration paradigm and an end-to-end procedure that avoids inefficient feature engineering and error propagation.

Our method is evaluated on the PDTB 2.0 benchmark in a variety of experimental settings. The proposed adversarial model greatly improves over standalone neural models and previous best-performing approaches. We also demonstrate that our implicit recognition network successfully imitates and extracts crucial hidden representations.

We begin by briefly reviewing related work in section 2. Section 3 presents the proposed adversarial model. Section 4 shows substantially improved experimental results over previous methods. Section 5 discusses extensions and future work.

2 Related Work

2.1 Implicit Discourse Relation Recognition

There has been a surge of interest in implicit discourse parsing since the release of PDTB (Prasad et al., 2008), the first large discourse corpus distinguishing implicit examples from explicit ones. A large set of work has focused on direct classification based on observed sentences, including structured methods with linguistically-informed features (Lin et al., 2009; Pitler et al., 2009; Zhou et al., 2010), end-to-end neural models (Qin et al., 2016b,c; Chen et al., 2016a; Liu and Li, 2016), and combined approaches (Ji and Eisenstein, 2015; Ji et al., 2016). However, the lacking of connective cues makes learning purely from contextual semantics full of challenges.

Prior work has attempted to leverage connective information. Zhou et al. (2010) also incorporate implicit connectives, but in a pipeline manner by first predicting the implicit connective with a language model and determining discourse relation accordingly. Instead of treating implicit connectives as intermediate prediction targets which can suffer from error propagation, we use the connectives to induce highly discriminative features to guide the learning of an implicit network, serving as an adaptive regularization mechanism for en-

hanced robustness and generalization. Our framework is also end-to-end, avoiding costly feature engineering. Another notable line aims at adapting explicit examples for data synthesis (Biran and McKeown, 2013; Rutherford and Xue, 2015; Braud and Denis, 2015; Ji et al., 2015), multi-task learning (Lan et al., 2013; Liu et al., 2016), and word representation (Braud and Denis, 2016). Our work is orthogonal and complementary to these methods, as we use implicit connectives which have been annotated for implicit examples.

2.2 Adversarial Networks

Deep neural networks have gained impressive success in various natural language processing tasks (Wang et al., 2016; Zhang et al., 2016b; Cai et al., 2017), in which adversarial networks have been shown especially effective in deep generative modeling (Goodfellow et al., 2014) and domain adaptation (Ganin et al., 2016). Generative adversarial nets (Goodfellow et al., 2014) learn to produce realistic samples through competition between a generator and a real/fake discriminator. Professor forcing (Lamb et al., 2016) applies a similar idea to improve long-term generation of a recurrent neural language model. Other approaches (Chen et al., 2016b; Hu et al., 2017; Liang et al., 2017) extend the framework for controllable image/text generation. Li et al. (2015); Salimans et al. (2016) propose feature matching which trains generators to match the statistics of real/fake examples. Their features are extracted by the discriminator rather than the classifier networks as in our case. Our work differs from the above since we consider the context of discriminative modeling. Adversarial domain adaptation forces a neural network to learn domain-invariant features using a classifier that distinguishes the domain of the network's input data based on the hidden feature. Our adversarial framework is distinct in that besides the implicit relation network we construct a second neural network serving as a teacher model for feature emulation.

To the best of our knowledge, this is the first to employ the idea of adversarial learning in the context of discourse parsing. We propose a novel connective exploiting scheme based on feature imitation, and to this end derive a new adversarial framework, achieving substantial performance gain over existing methods. The proposed approach is generally applicable to other tasks for utilizing any indicative side information. We give more discussions in section 5.

3 Adversarial Method

Discourse connectives are key indicators for discourse relation. In the annotation procedure of the PDTB implicit relation benchmark, annotators inserted implicit connective expressions between adjacent sentences to lexicalize abstract relations and help with final decisions. Our model aims at making full use of the provided implicit connectives at training time to regulate learning of implicit relation recognizer, encouraging extraction of highly discriminative semantics from raw arguments, and improving generalization at test time. Our method provides a novel adversarial framework that leverages connective information in a flexible adaptive manner, and is efficiently trained end-to-end through standard back-propagation.

The basic idea of the proposed approach is simple. We want our implicit relation recognizer, which predicts the underlying relation of sentence arguments without discourse connective, to have prediction behaviors close to a *connective-augmented* relation recognizer which is provided with a discourse connective in addition to the arguments. The connective-augmented recognizer is in analogy to an annotator with the help of connectives as in the human annotation process, and the implicit recognizer would be improved by learning from such an "informed" annotator. Specifically, we want the latent features extracted by the two models to match as closely as possible, which explicitly transfers the discriminability of the connective-augmented representations to implicit ones.

To this end, instead of manually selecting a closeness metric, we take advantage of the adversarial framework by constructing a two-player zero-sum game between the implicit recognizer and a rival discriminator. The discriminator attempts to distinguish between the features extracted by the two relation models, while the implicit relation model is trained to maximize the accuracy on implicit data, and at the same time to confuse the discriminator.

In the next we first present the overall architecture of the proposed approach (section 3.1), then develop the training procedure (section 3.2). The components are realized as deep (convolutional) neural networks, with detailed modeling choices
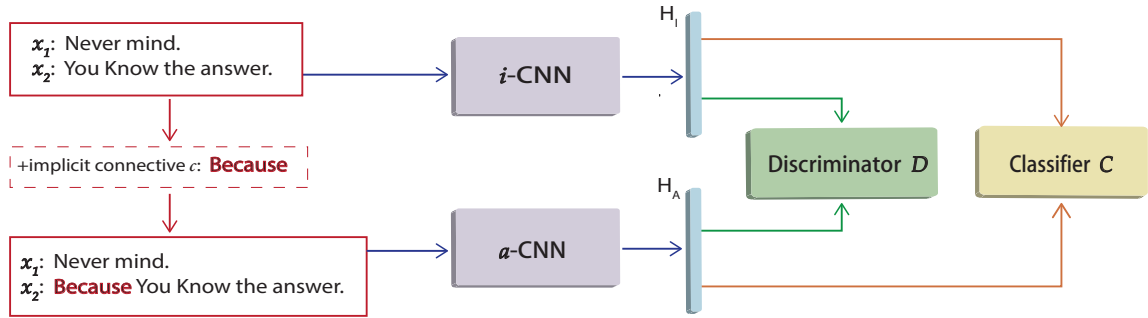
Figure 1: Architecture of the proposed method. The framework contains three main components: 1) an implicit relation network *i-CNN* over raw sentence arguments, 2) a connective-augmented relation network *a-CNN* whose inputs are augmented with implicit connectives, and 3) a discriminator distinguishing between the features by the two networks. The features are fed to the final classifier for relation classification. The discriminator and *i-CNN* form an adversarial pair for feature imitation. At test time, the implicit network *i-CNN* with the classifier is used for prediction.

discussed in section 3.3.

3.1 Model Architecture

Let $(\boldsymbol{x}, y)$ be a pair of input and output of implicit relation classification, where $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ is a pair of sentence arguments, and y is the underlying discourse relation. Each training example also includes an annotated implicit connective c that best expresses the relation. Figure 1 shows the architecture of our framework.

The neural model for implicit relation classification (*i-CNN* in the figure) extracts latent representation from the arguments, denoted as $H_I(\boldsymbol{x}_1, \boldsymbol{x}_2)$, and feeds the feature into a classifier C for final prediction $C(H_I(\boldsymbol{x}_1, \boldsymbol{x}_2))$. For ease of notation, we will also use $H_I(\boldsymbol{x})$ to denote the latent feature on data $\boldsymbol{x}$.

The second relation network (*a-CNN*) takes as inputs the sentence arguments along with an implicit connective, to induce the connective-augmented representation $H_A(\boldsymbol{x}_1, \boldsymbol{x}_2, c)$, and obtains relation prediction $C(H_A(\boldsymbol{x}_1, \boldsymbol{x}_2, c))$. Note that the same final classifier C is used for both networks, so that the feature representations by the two networks are ensured to be within the same semantic space, enabling feature emulation as presented shortly.

We further pair the implicit network with a rival discriminator D to form our adversarial game. The discriminator is to differentiate between the reasoning behaviors of the implicit network *i-CNN* and the augmented network *a-CNN*. Specifically, D is a binary classifier that takes as inputs a la-

tent feature H derived from either *i-CNN* or *a-CNN* given appropriate data (where implicit connectives is either missing or present, respectively). The output $D(H)$ estimates the probability that H comes from the connective-augmented *a-CNN* rather than *i-CNN*.

3.2 Training Procedure

The system is trained through an alternating optimization procedure that updates the components in an interleaved manner. In this section, we first present the training objective for each component, and then give the overall training algorithm.

Let $\boldsymbol{\theta}_D$ denote the parameters of the discriminator. The training objective of D is straightforward, i.e., to maximize the probability of correctly distinguishing the input features:

$$\max_{\boldsymbol{\theta}_D} \mathcal{L}_D = \mathbb{E}_{(\boldsymbol{x},c,y)\sim\text{data}} \Big[\log D(H_A(\boldsymbol{x}, c); \boldsymbol{\theta}_D) + \log(1 - D(H_I(\boldsymbol{x}); \boldsymbol{\theta}_D)) \Big], \quad (1)$$

where $\mathbb{E}_{(\boldsymbol{x},c,y)\sim\text{data}}[\cdot]$ denotes the expectation in terms of the data distribution.

We denote the parameters of the implicit network *i-CNN* and the classifier C as $\boldsymbol{\theta}_I$ and $\boldsymbol{\theta}_C$, respectively. The model is then trained to (a) correctly classify relations in training data and (b) produce salient features close to connective-augmented ones. The first objective can be fulfilled by minimizing the usual cross-entropy loss:

$$\mathcal{L}_{I,C}(\boldsymbol{\theta}_I, \boldsymbol{\theta}_C) = \mathbb{E}_{(\boldsymbol{x},y)\sim\text{data}} \Big[J\big(C(H_I(\boldsymbol{x}; \boldsymbol{\theta}_I); \boldsymbol{\theta}_C), y\big) \Big], \quad (2)$$

Algorithm 1 Adversarial Model for Implicit Recognition

Input: Training data $\{(\boldsymbol{x}, c, y)_n\}$

Parameters: λ_1, λ_2 – balancing parameters

1: Initialize $\{\boldsymbol{\theta}_I, \boldsymbol{\theta}_C\}$ and $\{\boldsymbol{\theta}_A\}$ by minimizing Eq.(2) and Eq.(4), respectively
2: **repeat**
3:     Train the discriminator through Eq.(1)
4:     Train the relation models through Eq.(5)
5: **until** convergence

Output: Adversarially enhanced implicit relation network *i-CNN* with classifier C for prediction

where $J(\boldsymbol{p}, y) = -\sum_k \mathbb{I}(y = k) \log p_k$ is the cross-entropy loss between predictive distribution $\boldsymbol{p}$ and ground-truth label y. We achieve objective (b) by minimizing the discriminator's chance of correctly telling apart the features:

$$\mathcal{L}_I(\boldsymbol{\theta}_I) = \mathbb{E}_{\boldsymbol{x} \sim \text{data}} \Big[\log \big(1 - D(H_I(\boldsymbol{x}; \boldsymbol{\theta}_I)) \big) \Big]. \quad (3)$$

The parameters of the augmented network *a-CNN*, denoted as $\boldsymbol{\theta}_A$, can be learned by simply fitting to the data, i.e., minimizing the cross-entropy loss as follows:

$$\mathcal{L}_A(\boldsymbol{\theta}_A) = \mathbb{E}_{(\boldsymbol{x}, c, y) \sim \text{data}} \Big[J \big(C(H_A(\boldsymbol{x}, c; \boldsymbol{\theta}_A)), y \big) \Big]. \quad (4)$$

As mentioned above, here we use the same classifier C as for the implicit network, forcing a unified feature space of both networks. We combine the above objectives Eqs.(2)-(4) of the relation classifiers and minimize the joint loss:

$$\min_{\boldsymbol{\theta}_I, \boldsymbol{\theta}_A, \boldsymbol{\theta}_C} \mathcal{L}_{I,A,C} = \mathcal{L}_{I,C}(\boldsymbol{\theta}_I, \boldsymbol{\theta}_C) + \lambda_1 \mathcal{L}_I(\boldsymbol{\theta}_I) + \lambda_2 \mathcal{L}_A(\boldsymbol{\theta}_A), \quad (5)$$

where λ_1 and λ_2 are two balancing parameters calibrating the weights of the classification losses and the feature-regulating loss. In practice, we pre-train the implicit and augmented networks independently by minimizing Eq.(2) and Eq.(4), respectively. In the adversarial training process, we found setting $\lambda_2 = 0$ gives stable convergence. That is, the connective-augmented features are fixed after the pre-training stage.

Algorithm 1 summarizes the training procedure, where we interleave the optimization of Eq.(1) and Eq.(5) at each iteration. More practical details are provided in section 4. We instantiate all modules as neural networks (section 3.3) which are differentiable, and perform the optimization efficiently through standard stochastic gradient descent and back-propagation.
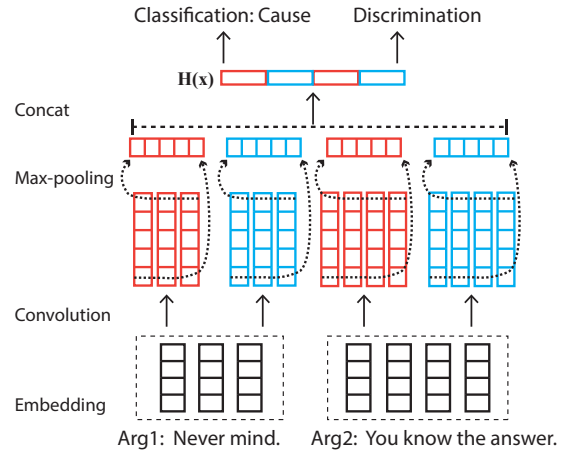

Figure 2: Neural structure of *i-CNN*. Two sets of convolutional filters are shown, with the corresponding features in red and blue, respectively. The weights of the filters on two input arguments are tied.

Through Eq.(1) and Eq.(3), the discriminator and the implicit relation network follow a minimax competition, which drives both to improve until the implicit feature representations are close to the connective-augmented latent representations, encouraging the implicit network to extract highly discriminative features from raw sentence arguments for relation classification. Alternatively, we can see Eq.(3) as an *adaptive* regularization on the implicit model, which, compared to pre-fixed regularizors such as ℓ_2-regularization, provides a more flexible, self-calibrated mechanism to improve generalization ability.
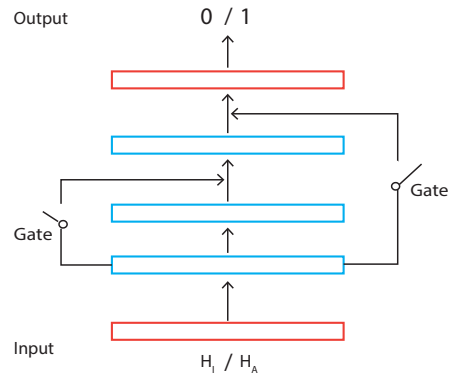

Figure 3: Neural structure of the discriminator D.

3.3 Component Structures

We have presented our adversarial framework for implicit relation classification. We now discuss the model realization of each component. All components of the framework are parameterized with neural networks. Distinct roles of the modules in the framework lead to different modeling choices.

Relation Classification Networks Figure 2 illustrates the structure of the implicit relation network *i-CNN*. We use a convolutional network as it is a common architectural choice for discourse parsing. The network takes as inputs the word vectors of the tokens in each sentence argument, and maps each argument to intermediate features through a shared convolutional layer. The resulting representations are then concatenated and fed into a max pooling layer to select most salient features as the final representation. The final classifier C is a simple fully-connected layer followed by a softmax classifier.

The connective-augmented network *a-CNN* has a similar structure as *i-CNN*, wherein implicit connective is appended to the second sentence as input. The key difference from *i-CNN* is that here we adopt *average k-max* pooling, which takes the average of the top-k maximum values in each pooling window. The reason is to prevent the network from solely selecting the connective induced features (which are typically the most salient features) which would be the case when using max pooling, but instead force it to also attend to contextual features derived from the arguments. This facilitates more homogeneous output features of the two networks, and thus facilitates feature imitation. In all the experiments we fixed $k = 2$.

Discriminator The discriminator is a binary classifier to identify the correct source of an input feature vector. To make it a strong rival to the feature imitating network (*i-CNN*), we model the discriminator as a multi-layer perceptron (MLP) enhanced with gated mechanism for efficient information flow (Srivastava et al., 2015; Qin et al., 2016c), as shown in Figure 3.

4 Experiments

We demonstrate the effectiveness of our approach both quantitatively and qualitatively with extensive experiments. We evaluate prediction performance on the PDTB benchmark in different settings. Our method substantially improves over a diverse set of previous models, especially in the practical multi-class classification task. We perform in-depth analysis of the model behaviors, and show our adversarial framework successfully enables the implicit relation model to imitate and learn discriminative features.

4.1 Experiment Setup

We use PDTB 2.0[1], one of the largest manually annotated discourse relation corpus. The dataset contains 16,224 implicit relation instances in total, with three levels of senses: Level-1 *Class*, Level-2 *Type*, and Level-3 *Subtypes*. The 1st level consists of four major relation *Class*es: COMPARISON, CONTINGENCY, EXPANSION and TEMPORAL. The 2nd level contains 16 *Type*s.

To make extensive comparison with prior work of implicit discourse relation classification, we evaluate on two popular experimental settings: 1) multi-class classification for 2nd-level types (Lin et al., 2009; Ji and Eisenstein, 2015), and 2) one-versus-others binary classifications for 1st-level classes (Pitler et al., 2009). We describe the detailed configurations in the following respective sections. We will focus our analysis on the multi-class classification setting, which is most realistic in practice and serves as a building block for a complete discourse parser such as that for the shared tasks of CoNLL-2015 and 2016 (Xue et al., 2015, 2016).

Model Training Here we provide the detailed architecture configurations of each component we used in the experiments.

- Throughout the experiments *i-CNN* and *a-CNN* contain 3 sets of convolutional filters with the filter sizes selected on the dev set. Table 1 lists the filter configurations of the convolutional layer in *i-CNN* and *a-CNN* in different tasks. As described in section 3.3, following the convolutional layer is a max pooling layer in *i-CNN*, and an average k-max pooling layer with $k = 2$ in *a-CNN*.

- The final single-layer classifier C contains 512 neurons with *tanh* activation function.

- The discriminator D consists of 4 fully-connected layers, with 2 gated pathways from layer 1 to layer 3 and layer 4 (see Figure 3).

---

[1] http://www.seas.upenn.edu/~pdtb/

Task	Filter sizes	Filter number
PDTB-Lin	2, 4, 8	3×256
PDTB-Ji	2, 5, 10	3×256
One-vs-all	2, 5, 10	3×1024

Table 1: The convolutional architectures of *i-CNN* and *a-CNN* in different tasks (section 4). For example, in PDTB-Lin, we use 3 sets of filters, each of which is of size 2, 4, and 8, respectively; and each set has 256 filters.

The size of each layer is set to 1024 and is fixed in all the experiments.

- We set the dimension of the input word vectors to 300 and initialize with pre-trained word2vec (Mikolov et al., 2013). The maximum length of sentence argument is set to 80. Truncation and zero-padding are applied when necessary.

All experiments were performed on a TITAN-X GPU and 128GB RAM, with neural implementation based on Tensorflow[2].

For adversarial model training, it is critical to keep balance between the progress of the two players. We use a simple strategy which at each iteration optimizes the discriminator and the implicit relation network on a randomly-sampled mini-batch. We found this is enough to stabilize the training. The neural parameters are trained using AdaGrad (Duchi et al., 2011) with an initial learning rate of 0.001. For the balancing parameters in Eq.(5), we set $\lambda_1 = 0.1$, while $\lambda_2 = 0$. That is, after the initialization stage the weights of the connective-augmented network *a-CNN* are fixed. This has been shown capable of giving stable and good predictive performance for our system.

4.2 Implicit Relation Classification

We will mainly focus on the general multi-class classification problem in two alternative settings adopted in prior work, showing the superiority of our model over previous state of the arts. We perform in-depth comparison with carefully designed baselines, providing empirical insights into the working mechanism of the proposed framework. For broader comparisons we also report the performance in the one-versus-all setting.

	Model	PDTB-Lin	PDTB-Ji
1	Word-vector	34.07	36.86
2	CNN	43.12	44.51
3	Ensemble	42.17	44.27
4	Multi-task	43.73	44.75
5	ℓ_2-reg	44.12	45.33
6	Lin et al. (2009)	40.20	-
7	Lin et al. (2009) +Brown clusters	-	40.66
8	Ji and Eisenstein (2015)	-	44.59
9	Qin et al. (2016a)	43.81	45.04
10	Ours	**44.65**	**46.23**

Table 2: Accuracy (%) on the test sets of the PDTB-Lin and PDTB-Ji settings for **multi-class classification**. Please see the text for more details.

Multi-class Classifications

We first adopt the standard PDTB splitting convention following (Lin et al., 2009), denoted as PDTB-Lin, where sections 2-21, 22, and 23 are used as training, dev, and test sets, respectively. The most frequent 11 types of relations are selected in the task. During training, instances with more than one annotated relation types are considered as multiple instances, each of which has one of the annotations. At test time, a prediction that matches one of the gold types is considered as correct. The test set contains 766 examples. More details are in (Lin et al., 2009). An alternative, slightly different multi-class setting is used in (Ji and Eisenstein, 2015), denoted as PDTB-Ji, where sections 2-20, 0-1, and 21-22 are used as training, dev, and test sets, respectively. The resulting test set contains 1039 examples. We also evaluate in this setting for thorough comparisons.

Table 2 shows the classification accuracy in both of the settings. We see that our model (Row 10) achieves state-of-the-art performance, greatly outperforming previous methods (Rows 6-9) with various modeling paradigms, including the linguistic feature-based model (Lin et al., 2009), pure neural methods (Qin et al., 2016c), and combined approach (Ji and Eisenstein, 2015).

To obtain better insights into the working mechanism of our method, we further compare with a set of carefully selected baselines as shown in Rows 1-5. 1) "Word-vector" sums over the word vectors for sentence representation, showing the base effect of word embeddings. 2) "CNN" is a standalone convolutional net having the exact same architecture with our implicit rela-

tion network. Our model trained within the proposed framework provides significant improvement, showing the benefits of utilizing implicit connectives at training time. 3) "Ensemble" has the same neural architecture with the proposed framework except that the input of *a-CNN* is not augmented with implicit connectives. This essentially is an ensemble of two implicit recognition networks. We see that the method performs even inferior to the single CNN model. This further confirms the necessity of exploiting connective information. 4) "Multi-task" is the convolutional net augmented with an additional task of simultaneously predicting the implicit connectives based on the network features. As a straightforward way of incorporating connectives, we see that the method slightly improves over the stand-alone CNN, while falling behind our approach with a large margin. This indicates that our proposed feature imitation is a more effective scheme for making use of implicit connectives. 5) At last, "ℓ_2-reg" also implements feature mimicking by imposing an ℓ_2 distance penalty between the implicit relation features and connective-augmented features. We see that the simple model has obtained improvement over previous best-performing systems in both settings, further validating the idea of imitation. However, in contrast to the fixed ℓ_2 regularization, our adversarial framework provides an adaptive mechanism, which is more flexible and performs better as shown in the table.

Model	COMP.	CONT.	EXP.	TEMP.
Pitler et al. (2009)	21.96	47.13	-	16.76
Qin et al. (2016c)	**41.55**	**57.32**	71.50	35.43
Zhang et al. (2016a)	35.88	50.56	71.48	29.54
Zhou et al. (2010)	31.79	47.16	70.11	20.30
Liu and Li (2016)	36.70	54.48	70.43	**38.84**
Chen et al. (2016a)	40.17	54.76	-	31.32
Ours	40.87	54.56	**72.38**	36.20

Table 3: Comparisons of F_1 scores (%) for binary classification.

One-versus-all Classifications

We also report the results of four one-versus-all binary classifications for more comparisons with prior work. We follow the conventional experimental setting (Pitler et al., 2009) by selecting sections 2-20, 21-22, and 0-1 as training, dev, and test sets. Table 4 lists the statistics of the data.

Following previous work, Table 3 reports the F1

Relation	Train	Dev	Test
Comparison	1942/1942	197/986	152/894
Contigency	3342/3342	295/888	279/767
Expansion	7004/7004	671/512	574/472
Temporal	760/760	64/1119	85/961

Table 4: Distributions of positive and negative instances from the train/dev/test sets in four binary relation classification tasks.

scores. Our method outperforms most of the prior systems in all the tasks. We achieve state-of-the-art performance in recognition of the Expansion relation, and obtain comparable scores with the best-performing methods in each of the other relations, respectively. Notably, our feature imitation scheme greatly improves over (Zhou et al., 2010) which leverages implicit connectives as an intermediate prediction task. This provides additional evidence for the effectiveness of our approach.

4.3 Qualitative Analysis

We now take a closer look into the modeling behavior of our framework, by investigating the process of the adversarial game during training, as well as the feature imitation effects.

Figure 4 demonstrates the training progress of different components. The *a-CNN* network keeps high predictive accuracy as implicit connectives are given, showing the importance of connective cues. The rise-and-fall patterns in the accuracy of the discriminator clearly show its competition with the implicit relation network *i-CNN* as training goes. At first few iterations the accuracy of the discriminator increases quickly to over 0.9, while at late stage the accuracy drops to around 0.6, showing that the discriminator is getting confused by *i-CNN* (an accuracy of 0.5 indicates full confusion). The *i-CNN* network keeps improving in terms of implicit relation classification accuracy, as it is gradually fitting to the data and simultaneously learning increasingly discriminative features by mimicking *a-CNN*. The system exhibits similar learning patterns in the two different settings, showing the stability of the training strategy.

We finally visualize the output feature vectors of *i-CNN* and *a-CNN* using the t-SNE method (Maaten and Hinton, 2008) in Figure 5. Without feature imitation, the extracted features by the two networks are clearly separated (Figure 5(a)). In contrast, as shown in Figures 5(b)-(c), the feature vectors are increasingly mixed as training proceeds. Thus our framework has suc-
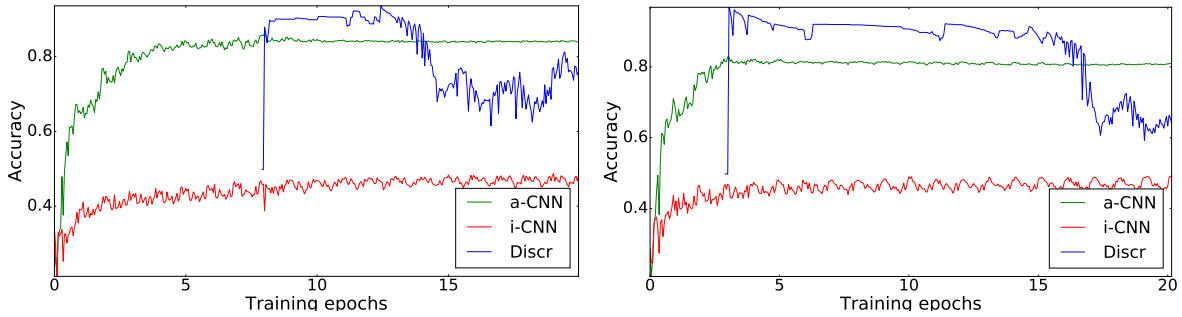
Figure 4: (Best viewed in colors.) Test-set performance of three components over training epochs. Relation networks *a-CNN* and *i-CNN* are measured with multi-class classification accuracy (with or without implicit connectives, respectively), while the discriminator is evaluated with binary classification accuracy. **Top**: the PDTB-Lin setting (Lin et al., 2009), where first 8 epochs are for initialization stage (thus the discriminator is fixed and not shown); **Bottom**: the PDTB-Ji setting (Ji and Eisenstein, 2015), where first 3 epochs are for initialization.
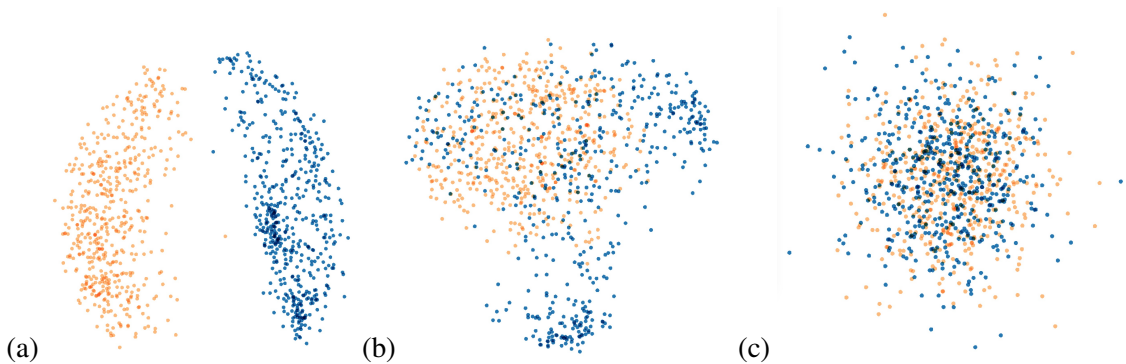

(a)   (b)   (c)

Figure 5: (Best viewed in colors.) Visualizations of the extracted hidden features by the implicit relation network *i-CNN* (blue) and connective-augmented relation network *a-CNN* (orange), in the multi-class classification setting (Lin et al., 2009). **(a)** Two networks are trained without adversary (with shared classifier); **(b)** Two networks are trained within our framework at epoch 10; **(c)** at epoch 20. The implicit relation network successfully imitates the connective-augmented features through the adversarial game. Visualization is conducted with the t-SNE algorithm (Maaten and Hinton, 2008).

cessfully driven *i-CNN* to induce similar representations with *a-CNN*, even though connectives are not present.

5 Discussions

We have developed an adversarial neural framework that facilitates an implicit relation network to extract highly discriminative features by mimicking a connective-augmented network. Our method achieved state-of-the-art performance for implicit discourse relation classification. Besides implicit connective examples, our model can naturally exploit enormous explicit connective data to further improve discourse parsing.

The proposed adversarial feature imitation scheme is also generally applicable to other context to incorporate indicative side information

available at training time for enhanced inference. Our framework shares a similar spirit of the iterative knowledge distillation method (Hu et al., 2016a,b) which train a "student" network to mimic the classification behavior of a knowledge-informed "teacher" network. Our approach encourages imitation on the feature level instead of the final prediction level. This allows our approach to apply to regression tasks, and more interestingly, the context in which the student and teacher networks have different prediction outputs, e.g., performing different tasks, while transferring knowledge between each other can be beneficial. Besides, our adversarial mechanism provides an adaptive metric to measure and drive the imitation procedure.

References

Or Biran and Kathleen McKeown. 2013. Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL, Volume 2: Short Papers)*. Sofia, Bulgaria, pages 69–73.

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)*. Lisbon, Portugal, pages 2201–2211.

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 (EMNLP)*. Austin, Texas, pages 203–213.

Deng Cai, Hai Zhao, Zhisong Zhang, Yuan Xin, Yongjian Wu, and Feiyue Huang. 2017. Fast and accurate neural word segmentation for Chinese. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*. Vancouver, Canada.

Changge Chen, Peilu Wang, and Hai Zhao. 2015. Shallow discourse parsing using constituent parsing tree. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task (CONLL)*. Beijing, China, pages 37–41.

Changge Chen and Hai Zhao. 2015. Deceptive opinion spam detection using deep level linguistic feature. In *The 4th CCF Conference on Natural Language Processing and Chinese Computing (NLPCC 2015), LNCS*. Nanchang, China, volume 9362, pages 465–474.

Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016a. 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 Volume 1: Long Papers)*. Berlin, Germany, pages 1726–1735.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016b. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*. pages 2172–2180.

Xilun Chen, Ben Athiwaratkun, Yu Sun, Kilian Weinberger, and Claire Cardie. 2016c. Adversarial deep averaging networks for cross-lingual sentiment classification. *arXiv preprint arXiv:1606.01614* .

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

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research* 17(59):1–35.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, T. Raymond 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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. pages 2672–2680.

Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric P Xing. 2016a. Harnessing deep neural networks with logic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*. Berlin, Germany, pages 2410–2420.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Controllable text generation. *arXiv preprint arXiv:1703.00955* .

Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric P Xing. 2016b. Deep neural networks with massive learned knowledge. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Austin, USA, pages 1670–1679.

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 (TACL)* 3:329–344.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse-driven language models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*. San Diego, California, pages 332–342.

Yangfeng Ji, Gongbo Zhang, and Jacob Eisenstein. 2015. Closing the gap: Domain adaptation from explicit to implicit discourse relations. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Lisbon, Portugal, pages 2219–2224.

Alex M Lamb, Anirudh Goyal, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. 2016. Professor forcing: A new algorithm for training recurrent networks. In *Advances In Neural Information Processing Systems*. pages 4601–4609.

Man Lan, Yu Xu, and Zhengyu Niu. 2013. 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 (ACL, Volume 1: Long Papers)*. Sofia, Bulgaria, pages 476–485.

Junyi Jessy Li, Marine Carpuat, and Ani Nenkova. 2014. Assessing the discourse factors that influence the quality of machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL, Volume 2: Short Papers)*. Baltimore, Maryland, pages 283–288.

Yujia Li, Kevin Swersky, and Richard S Zemel. 2015. Generative moment matching networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*. Lille, France, pages 1718–1727.

Zhongyi Li, Hai Zhao, Chenxi Pang, Lili Wang, and Huan Wang. 2016. A constituent syntactic parse tree based discourse parser. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning - Shared Task (CONLL)*. Berlin, Germany, pages 60–64.

Maria Liakata, Simon Dobnik, Shyamasree Saha, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2013. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Seattle, Washington, USA, pages 747–757.

Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. 2017. Recurrent topic-transition GAN for visual paragraph generation. *arXiv preprint arXiv:1703.07022* .

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 (EMNLP)*. Singapore, pages 343–351.

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)*. Austin, Texas, pages 1224–1233.

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

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9:2579–2605.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems (3)*. South Lake Tahoe, Nevada, USA, pages 3111–3119.

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 he Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*. Suntec, Singapore, pages 683–691.

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 *The sixth international conference on Language Resources and Evaluation (LREC)*. Marrakech, Morocco, pages 2961–2968.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016a. Implicit discourse relation recognition with context-aware character-enhanced embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan, pages 1914–1924.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016b. Shallow discourse parsing using convolutional neural network. In *Proceedings of the CoNLL-16 shared task*. Berlin, Germany, pages 70–77.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016c. 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, Texas, pages 2263–2270.

Attapol Rutherford and Nianwen Xue. 2015. 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 (NAACL: HLT)*. Denver, Colorado, pages 799–808.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*. pages 2226–2234.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Highway networks. *arXiv preprint arXiv:1505.00387* .

Peilu Wang, Yao Qian, Frank K. Soong, Lei He, and Hai Zhao. 2016. Learning distributed word representations for bidirectional LSTM recurrent neural network. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. San Diego, California, pages 527–533.

Yu Xu, Man Lan, Yue Lu, Zheng Yu Niu, and Chew Lim Tan. 2012. Connective prediction using machine learning for implicit discourse relation classification. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*. Brisbane, Australia, pages 1–8.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning - Shared Task (CoNLL)*. Beijing, China, pages 1–16.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Bonnie Webber, Attapol Rutherford, Chuan Wang, and Hongmin Wang. 2016. The CoNLL-2016 shared task on shallow discourse parsing. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning - Shared Task (CoNLL)*. Berlin, Germany, pages 1–19.

Biao Zhang, Deyi Xiong, jinsong su, Qun Liu, Rongrong Ji, Hong Duan, and Min Zhang. 2016a. Variational neural discourse relation recognizer. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Austin, Texas, pages 382–391.

Zhisong Zhang, Hai Zhao, and Lianhui Qin. 2016b. Probabilistic graph-based dependency parsing with convolutional neural network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*. Berlin, Germany, pages 1382–1392.

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 (CoLING 2010)*. Beijing, China, pages 1507–1514.