Improving Implicit Discourse Relation Classification by Modeling Inter-dependencies of Discourse Units in a Paragraph

Zeyu Dai, Ruihong Huang
Department of Computer Science and Engineering
Texas A&M University
{jzdaizeyu, huangrh}@tamu.edu

Abstract

We argue that semantic meanings of a sentence or clause can not be interpreted independently from the rest of a paragraph, or independently from all discourse relations and the overall paragraph-level discourse structure. With the goal of improving implicit discourse relation classification, we introduce a paragraph-level neural networks that model inter-dependencies between discourse units as well as discourse relation continuity and patterns, and predict a sequence of discourse relations in a paragraph. Experimental results show that our model outperforms the previous state-of-the-art systems on the benchmark corpus of PDTB.

1 Introduction

PDTB-style discourse relations, mostly defined between two adjacent text spans (i.e., discourse units, either clauses or sentences), specify how two discourse units are logically connected (e.g., causal, contrast). Recognizing discourse relations is one crucial step in discourse analysis and can be beneficial for many downstream NLP applications such as information extraction, machine translation and natural language generation.

Commonly, explicit discourse relations were distinguished from implicit ones, depending on whether a discourse connective (e.g., "because" and "after") appears between two discourse units (Prasad et al., 2008a). While explicit discourse relation detection can be framed as a discourse connective disambiguation problem (Pitler and Nenkova, 2009; Lin et al., 2014) and has achieved reasonable performance (F1 score > 90%), implicit discourse relations have no discourse connective and are especially difficult to identify (Lin et al., 2009, 2014; Xue et al., 2015). To fill the gap, implicit discourse relation prediction has drawn significant research interest recently and progress has been made (Chen et al., 2016; Liu and Li, 2016) by modeling compositional meanings of two discourse units and exploiting word interactions between discourse units using neural tensor networks or attention mechanisms in neural nets. However, most of existing approaches ignore wider paragraph-level contexts beyond the two discourse units that are examined for predicting a discourse relation in between.

To further improve implicit discourse relation prediction, we aim to improve discourse unit representations by positioning a discourse unit (DU) in its wider context of a paragraph. The key observation is that semantic meaning of a DU can not be interpreted independently from the rest of the paragraph that contains it, or independently from the overall paragraph-level discourse structure that involve the DU. Considering the following paragraph with four discourse relations, one relation between each two adjacent DUs:

(1): *[The Butler, Wis., manufacturer went public at $15.75 a share in August 1987,]$_{DU1}$* and ***(Explicit-Expansion)*** *[Mr. Sim's goal then was a $29 per-share price by 1992.]$_{DU2}$* ***(Implicit-Expansion)*** *[Strong earnings growth helped achieve that price far ahead of schedule, in August 1988.]$_{DU3}$* ***(Implicit-Comparison)*** *[The stock has since softened, trading around $25 a share last week and closing yesterday at $23 in national over-the-counter trading.]$_{DU4}$* But ***(Explicit-Comparison)*** *[Mr. Sim has set a fresh target of $50 a share by the end of reaching that goal.]$_{DU5}$*

Clearly, each DU is an integral part of the paragraph and not independent from other units. *First*, predicting a discourse relation may require understanding wider paragraph-level contexts beyond two relevant DUs and the overall discourse structure of a paragraph. For example, the implicit "Comparison" discourse relation between DU3 and DU4 is difficult to identify without the back-

Proceedings of NAACL-HLT 2018, pages 141–151
New Orleans, Louisiana, June 1 - 6, 2018. ©2018 Association for Computational Linguistics

ground information (the history of per-share price) introduced in DU1 and DU2. *Second*, a DU may be involved in multiple discourse relations (e.g., DU4 is connected with both DU3 and DU5 with a "Comparison" relation), therefore the pragmatic meaning representation of a DU should reflect all the discourse relations the unit was involved in. *Third*, implicit discourse relation prediction should benefit from modeling discourse relation continuity and patterns in a paragraph that involve easy-to-identify explicit discourse relations (e.g., "Implicit-Comparison" relation is followed by "Explicit-Comparison" in the above example).

Following these observations, we construct a neural net model to process a paragraph each time and jointly build meaning representations for all DUs in the paragraph. The learned DU representations are used to predict a sequence of discourse relations in the paragraph, including both implicit and explicit relations. Although explicit relations are not our focus, predicting an explicit relation will help to reveal the pragmatic roles of its two DUs and reconstruct their representations, which will facilitate predicting neighboring implicit discourse relations that involve one of the DUs.

In addition, we introduce two novel designs to further improve discourse relation classification performance of our paragraph-level neural net model. First, previous work has indicated that recognizing explicit and implicit discourse relations requires different strategies, we therefore untie parameters in the discourse relation prediction layer of the neural networks and train two separate classifiers for predicting explicit and implicit discourse relations respectively. This unique design has improved both implicit and explicit discourse relation identification performance. Second, we add a CRF layer on top of the discourse relation prediction layer to fine-tune a sequence of predicted discourse relations by modeling discourse relation continuity and patterns in a paragraph.

Experimental results show that the intuitive paragraph-level discourse relation prediction model achieves improved performance on PDTB for both implicit discourse relation classification and explicit discourse relation classification.

2 Related Work

2.1 Implicit Discourse Relation Recognition

Since the PDTB (Prasad et al., 2008b) corpus was created, a surge of studies (Pitler et al., 2009; Lin et al., 2009; Liu et al., 2016; Rutherford and Xue, 2016) have been conducted for predicting discourse relations, primarily focusing on the challenging task of implicit discourse relation classification when no explicit discourse connective phrase was presented. Early studies (Pitler et al., 2008; Lin et al., 2009, 2014; Rutherford and Xue, 2015) focused on extracting linguistic and semantic features from two discourse units. Recent research (Zhang et al., 2015; Rutherford et al., 2016; Ji and Eisenstein, 2015; Ji et al., 2016) tried to model compositional meanings of two discourse units by exploiting interactions between words in two units with more and more complicated neural network models, including the ones using neural tensor (Chen et al., 2016; Qin et al., 2016; Lei et al., 2017) and attention mechanisms (Liu and Li, 2016; Lan et al., 2017; Zhou et al., 2016). Another trend is to alleviate the shortage of annotated data by leveraging related external data, such as explicit discourse relations in PDTB (Liu et al., 2016; Lan et al., 2017; Qin et al., 2017) and unlabeled data obtained elsewhere (Rutherford and Xue, 2015; Lan et al., 2017), often in a multi-task joint learning framework.

However, nearly all the previous works assume that a pair of discourse units is independent from its wider paragraph-level contexts and build their discourse relation prediction models based on *only* two relevant discourse units. In contrast, we model inter-dependencies of discourse units in a paragraph when building discourse unit representations; in addition, we model global continuity and patterns in a sequence of discourse relations, including both implicit and explicit relations.

Hierarchical neural network models (Liu and Lapata, 2017; Li et al., 2016) have been applied to RST-style discourse parsing (Carlson et al., 2003) mainly for the purpose of generating text-level hierarchical discourse structures. In contrast, we use hierarchical neural network models to build context-aware sentence representations in order to improve implicit discourse relation prediction.

2.2 Paragraph Encoding

Abstracting latent representations from a long sequence of words, such as a paragraph, is a challenging task. While several novel neural network models (Zhang et al., 2017b,a) have been introduced in recent years for encoding a paragraph, Recurrent Neural Network (RNN)-based methods remain the most effective approaches. RNNs, especially the long-short term memory (LSTM) (Hochreiter and Schmidhuber, 1997) models, have been widely used to encode a paragraph for machine translation (Sutskever et al., 2014), dialogue systems (Serban et al., 2016) and text summarization (Nallapati et al., 2016) because of its ability in modeling long-distance dependencies between words. In addition, among four typical pooling methods (sum, mean, last and max) for calculating sentence representations from RNN-encoded hidden states for individual words, max-pooling along with bidirectional LSTM (Bi-LSTM) (Schuster and Paliwal, 1997) yields the current best universal sentence representation method (Conneau et al., 2017). We adopted a similar neural network architecture for paragraph encoding.

3 The Neural Network Model for Paragraph-level Discourse Relation Recognition

3.1 The Basic Model Architecture

Figure 1 illustrates the overall architecture of the discourse-level neural network model that consists of two Bi-LSTM layers, one max-pooling layer in between and one softmax prediction layer. The input of the neural network model is a paragraph containing a sequence of discourse units, while the output is a sequence of discourse relations with one relation between each pair of adjacent discourse units[1].

Given the words sequence of one paragraph as input, the lower Bi-LSTM layer will read the whole paragraph and calculate hidden states as word representations, and a max-pooling layer will be applied to abstract the representation of each discourse unit based on individual word representations. Then another Bi-LSTM layer will run over the sequence of discourse unit representations and compute new representations by further modeling semantic dependencies between discourse units within paragraph. The final softmax prediction layer will concatenate representations of two adjacent discourse units and predict the discourse relation between them.

Word Vectors as Input: The input of the paragraph-level discourse relation prediction model is a sequence of word vectors, one vector per word in the paragraph. In this work, we used the pre-trained 300-dimension Google English word2vec embeddings[2]. For each word that is not in the vocabulary of Google word2vec, we will randomly initialize a vector with each dimension sampled from the range $[-0.25, 0.25]$. In addition, recognizing key entities and discourse connective phrases is important for discourse relation recognition, therefore, we concatenate the raw word embeddings with extra linguistic features, specifically one-hot Part-Of-Speech tag embeddings and one-hot named entity tag embeddings[3].

Building Discourse Unit Representations: We aim to build discourse unit (DU) representations that sufficiently leverage cues for discourse relation prediction from paragraph-wide contexts, including the preceding and following discourse units in a paragraph. To process long paragraph-wide contexts, we take a bottom-up two-level abstraction approach and progressively generate a compositional representation of each word first (low level) and then generate a compositional representation of each discourse unit (high level), with a max-pooling operation in between. At both word-level and DU-level, we choose Bi-LSTM as our basic component for generating compositional representations, mainly considering its capability to capture long-distance dependencies between words (discourse units) and to incorporate influences of context words (discourse units) in each side.

Given a variable-length words sequence $X = (x_1, x_2, ..., x_L)$ in a paragraph, the word-level Bi-LSTM will process the input sequence by using two separate LSTMs, one process the word sequence from the left to right while the other follows the reversed direction. Therefore, at each word position t, we obtain two hidden states $\overrightarrow{h_t}, \overleftarrow{h_t}$. We concatenate them to get the word representation $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$. Then we apply max-pooling over the sequence of word representations for words in a discourse unit in order to get the discourse unit embedding:

[1] In PDTB, most of discourse relations were annotated between two adjacent sentences or two adjacent clauses. For exceptional cases, we applied heuristics to convert them.

[2] Downloaded from https://docs.google.com/uc?id=0B7XkCwpI5KDYNlNUTTlSS21pQmM

[3] Our feature-rich word embeddings are of dimension 343, including 300 dimensions for word2vec embeddings + 36 dimensions for Part-Of-Speech (POS) tags + 7 dimensions for named entity tags. We used the Stanford CoreNLP to generate POS tags and named entity tags.

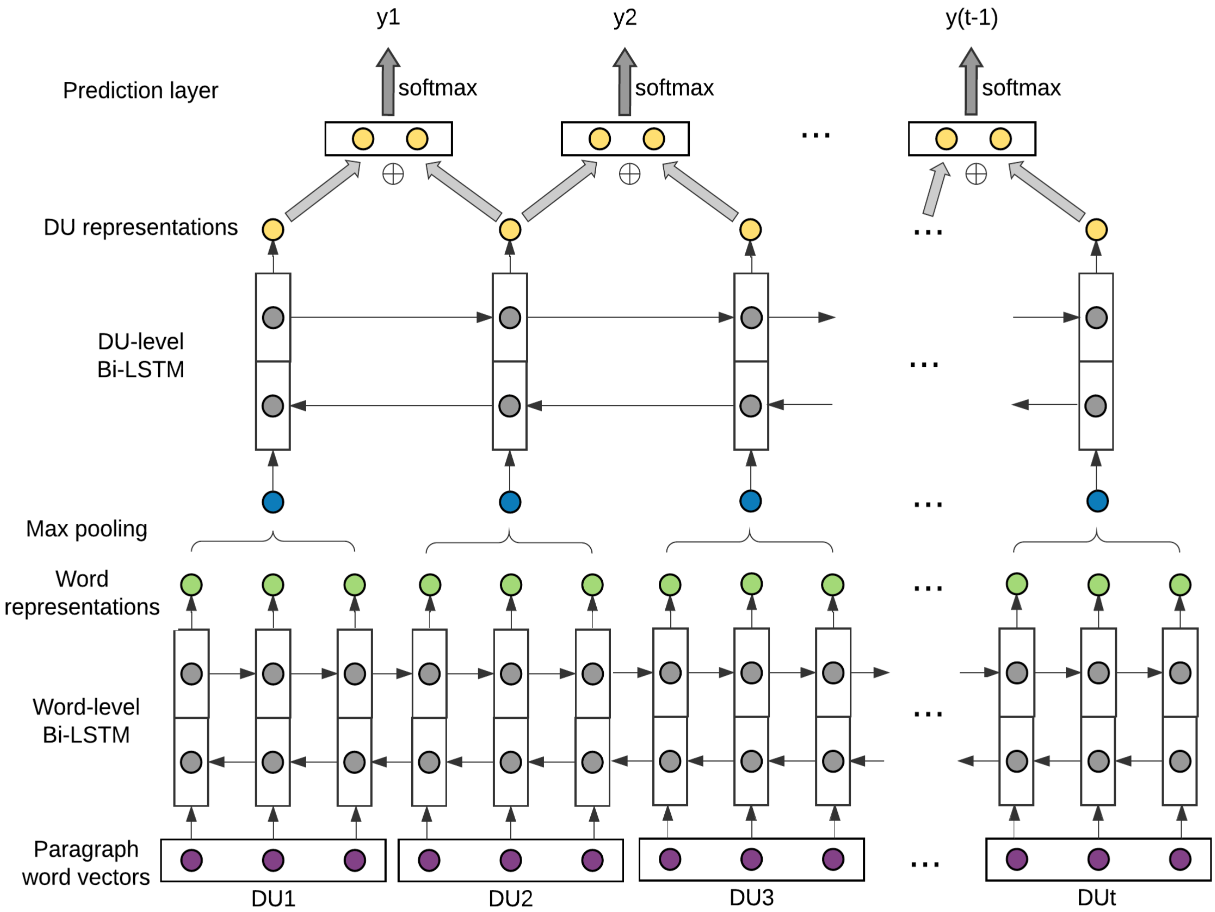

Figure 1: The Basic Model Architecture for Paragraph-level Discourse Relations Sequence Prediction.

$$MP_{DU}[j] = \max_{i=DU\_start}^{DU\_end} h_i[j] \tag{1}$$

$$where, 1 \leq j \leq hidden\_node\_size \tag{2}$$

Next, the DU-level Bi-LSTM will process the sequence of discourse unit embeddings in a paragraph and generate two hidden states $\overrightarrow{hDU_t}$ and $\overleftarrow{hDU_t}$ at each discourse unit position. We concatenate them to get the discourse unit representation $hDU_t = [\overrightarrow{hDU_t}, \overleftarrow{hDU_t}]$.

The Softmax Prediction Layer: Finally, we concatenate two adjacent discourse unit representations hDU_{t-1} and hDU_t and predict the discourse relation between them using a softmax function:

$$y_{t-1} = softmax(W_y * [hDU_{t-1}, hDU_t] + b_y) \tag{3}$$

3.2 Untie Parameters in the Softmax Prediction Layer (Implicit vs. Explicit)

Previous work (Pitler and Nenkova, 2009; Lin et al., 2014; Rutherford and Xue, 2016) has revealed that recognizing explicit vs. implicit discourse relations requires different strategies. Note that in the PDTB dataset, explicit discourse relations were distinguished from implicit ones, depending on whether a discourse connective exists between two discourse units. Therefore, explicit discourse relation detection can be simplified as a discourse connective phrase disambiguation problem (Pitler and Nenkova, 2009; Lin et al., 2014). On the contrary, predicting an implicit discourse relation should rely on understanding the overall

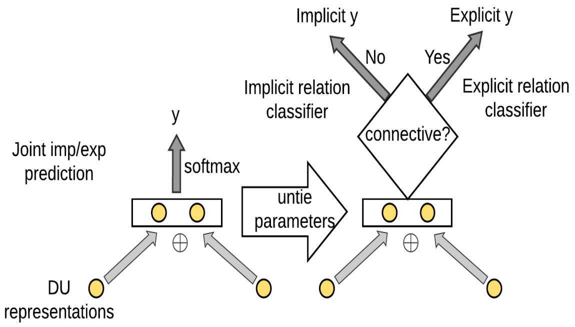

Figure 2: Untie Parameters in the Prediction Layer

contents of its two discourse units (Lin et al., 2014; Rutherford and Xue, 2016).

Considering the different natures of explicit vs. implicit discourse relation prediction, we decide to untie parameters at the final discourse relation prediction layer and train two softmax classifiers, as illustrated in Figure 2. The two classifiers have different sets of parameters, with one classifier for *only* implicit discourse relations and the other for *only* explicit discourse relations.

$$y_{t-1} = \begin{cases} softmax(W_{exp}[hDU_{t-1}, hDU_t] + b_{exp}), & exp \\ softmax(W_{imp}[hDU_{t-1}, hDU_t] + b_{imp}), & imp \end{cases} \tag{4}$$

The loss function used for the neural network training considers loss induced by both implicit relation prediction and explicit relation prediction:

$$Loss = Loss_{imp} + \alpha * Loss_{exp} \tag{5}$$

The α, in the full system, is set to be 1, which means that minimizing the loss in predicting either type of discourse relations is equally important. In the evaluation, we will also evaluate a system variant, where we will set $\alpha = 0$, which means that the neural network will not attempt to predict explicit discourse relations and implicit discourse relation prediction will not be influenced by predicting neighboring explicit discourse relations.

3.3 Fine-tune Discourse Relation Predictions Using a CRF Layer

Data analysis and many linguistic studies (Pitler et al., 2008; Asr and Demberg, 2012; Lascarides and Asher, 1993; Hobbs, 1985) have repeatedly shown that discourse relations feature continuity and patterns (e.g., a temporal relation is likely to be followed by another temporal relation). Especially, Pitler et al. (2008) firstly reported that patterns exist between implicit discourse relations and their neighboring explicit discourse relations.

Motivated by these observations, we aim to improve implicit discourse relation detection by making use of easily identifiable explicit discourse relations and taking into account global patterns of discourse relation distributions. Specifically, we add an extra CRF layer at the top of the softmax prediction layer (shown in figure 3) to fine-tune predicted discourse relations by considering their inter-dependencies.

The Conditional Random Fields (Lafferty et al., 2001) (CRF) layer updates a state transition matrix, which can effectively adjust the current label depending on proceeding and following labels. Both training and decoding of the CRF layer can be solved efficiently by using the Viterbi algorithm. With the CRF layer, the model jointly assigns a sequence of discourse relations between each two adjacent discourse units in a paragraph, including both implicit and explicit relations, by considering relevant discourse unit representations as well as global discourse relation patterns.

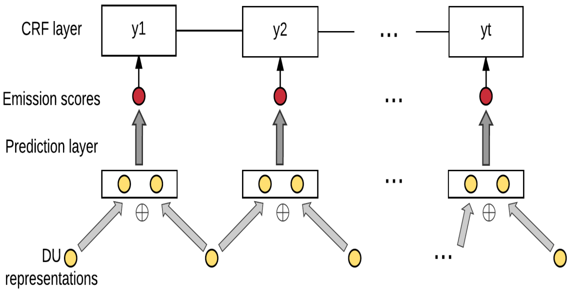

Figure 3: Fine-tune Discourse Relations with a CRF layer.

4 Evaluation

4.1 Dataset and Preprocessing

The Penn Discourse Treebank (PDTB): We experimented with PDTB v2.0 (Prasad et al., 2008b) which is the largest annotated corpus containing 36k discourse relations in 2,159 Wall Street Journal (WSJ) articles. In this work, we focus on the top-level[4] discourse relation senses which are consist of four major semantic classes: Comparison (Comp), Contingency (Cont), Expansion (Exp) and Temporal (Temp). We followed the same PDTB section partition (Rutherford and Xue, 2015) as previous work and used sections 2-20 as training set, sections 21-22 as test set, and sections 0-1 as development set. Table 1 presents the data distributions we collected from PDTB.

Preprocessing: The PDTB dataset documents its annotations as a list of discourse relations, with each relation associated with its two discourse units. To recover the paragraph context for a discourse relation, we match contents of its two annotated discourse units with all paragraphs in corresponding raw WSJ article. When all the matching was completed, each paragraph was split into a sequence of discourse units, with one discourse relation (implicit or explicit) between each two ad-

[4] In PDTB, the sense label of discourse relation was annotated hierarchically with three levels.

Type	Class	Train	Dev	Test	Total
Implicit	Comp	1942	197	152	2291
	Cont	3339	292	279	3910
	Exp	7003	671	574	8248
	Temp	760	64	85	909
Explicit	Comp	4184	422	364	4970
	Cont	2837	286	213	3336
	Exp	4612	481	424	5517
	Temp	2742	254	297	3293

Table 1: Distributions of Four Top-level Discourse Relations in PDTB.

# of DUs	2	3	4	5	>5
ratio	44%	25%	15%	7.3%	8.7%

Table 2: Distributions of Paragraphs.

jacent discourse units[5]. Following this method, we obtained 14,309 paragraphs in total, each contains 3.2 discourse units on average. Table 2 shows the distribution of paragraphs based on the number of discourse units in a paragraph.

4.2 Parameter Settings and Model Training

We tuned the parameters based on the best performance on the development set. We fixed the weights of word embeddings during training. All the LSTMs in our neural network use the hidden state size of 300. To avoid overfitting, we applied dropout (Hinton et al., 2012) with dropout ratio of 0.5 to both input and output of LSTM layers. To prevent the exploding gradient problem in training LSTMs, we adopt gradient clipping with gradient L2-norm threshold of 5.0. These parameters remain the same for all our proposed models as well as our own baseline models.

We chose the standard cross-entropy loss function for training our neural network model and adopted Adam (Kingma and Ba, 2014) optimizer with the initial learning rate of 5e-4 and a mini-batch size of 128[6]. If one instance is annotated with two labels (4% of all instances), we use both of them in loss calculation and regard the prediction as correct if model predicts one of the annotated labels. All the proposed models were implemented with Pytorch[7] and converged to the best performance within 20-40 epochs.

To alleviate the influence of randomness in neural network model training and obtain stable experimental results, we ran each of the proposed models and our own baseline models ten times and report the average performance of each model instead of the best performance as reported in many previous works.

4.3 Baseline Models and Systems

We compare the performance of our neural network model with several recent discourse relation recognition systems that only consider two relevant discourse units.

- (Rutherford and Xue, 2015): improves implicit discourse relation prediction by creating more training instances from the Gigaword corpus utilizing explicitly mentioned discourse connective phrases.
- (Chen et al., 2016): a gated relevance network (GRN) model with tensors to capture semantic interactions between words from two discourse units.
- (Liu et al., 2016): a convolutional neural network model that leverages relations between different styles of discourse relations annotations (PDTB and RST (Carlson et al., 2003)) in a multi-task joint learning framework.
- (Liu and Li, 2016): a multi-level attention-over-attention model to dynamically exploit features from two discourse units for recognizing an implicit discourse relation.
- (Qin et al., 2017): a novel pipelined adversarial framework to enable an adaptive imitation competition between the implicit network and a rival feature discriminator with access to connectives.
- (Lei et al., 2017): a Simple Word Interaction Model (SWIM) with tensors that captures both linear and quadratic relations between words from two discourse units.
- (Lan et al., 2017): an attention-based LSTM neural network that leverages explicit discourse relations in PDTB and unannotated external data in a multi-task joint learning framework.

[5] In several hundred discourse relations, one discourse unit is complex and can be further separated into two elementary discourse units, which can be illustrated as [DU1-DU2]-DU3. We simplify such cases to be a relation between DU2 and DU3.

[6] Counted as the number of discourse relations rather than paragraph instances.

[7] http://pytorch.org/

	Implicit						Explicit	
Model	Macro	Acc	Comp	Cont	Exp	Temp	Macro	Acc
(Rutherford and Xue, 2015)	40.50	57.10	-	-	-	-	-	-
(Liu et al., 2016)	44.98	57.27	-	-	-	-	-	-
(Liu and Li, 2016)	46.29	57.57	-	-	-	-	-	-
(Lei et al., 2017)	46.46	-	-	-	-	-	-	-
(Lan et al., 2017)	47.80	57.39	-	-	-	-	-	-
DU-pair level Discourse Relation Recognition (Our Own Baselines)								
Bi-LSTM	40.01	53.50	30.52	42.06	65.52	21.96	-	-
+ tensors	45.36	57.18	36.88	44.85	68.70	30.74	-	-
Paragraph level Discourse Relation Recognition								
Basic System Variant ($\alpha = 0$)	47.56	56.88	37.12	46.47	67.72	38.92	-	-
Basic System ($\alpha = 1$)	48.10	57.52	37.33	47.89	68.39	38.80	91.93	92.89
+ Untie Parameters	48.69	**58.20**	37.68	49.19	**68.86**	39.04	**93.70**	**94.46**
+ the CRF Layer	**48.82**	57.44	**37.72**	**49.39**	67.45	**40.70**	93.21	93.98

Table 3: Multi-class Classification Results on PDTB. We report accuracy (Acc) and macro-average F1-scores for both explicit and implicit discourse relation predictions. We also report class-wise F1 scores.

4.4 Evaluation Settings

On the PDTB corpus, both binary classification and multi-way classification settings are commonly used to evaluate the implicit discourse relation recognition performance. We noticed that all the recent works report class-wise implicit relation prediction performance in the binary classification setting, while none of them report detailed performance in the multi-way classification setting. In the binary classification setting, separate "one-versus-all" binary classifiers were trained, and each classifier is to identify one class of discourse relations. Although separate classifiers are generally more flexible in combating with imbalanced distributions of discourse relation classes and obtain higher class-wise prediction performance, one pair of discourse units may be tagged with all four discourse relations without proper conflict resolution. Therefore, the multi-way classification setting is more appropriate and natural in evaluating a practical end-to-end discourse parser, and we mainly evaluate our proposed models using the four-way multi-class classification setting.

Since none of the recent previous work reported class-wise implicit relation classification performance in the multi-way classification setting, for better comparisons, we re-implemented the neural tensor network architecture (so-called SWIM in (Lei et al., 2017)) which is essentially a Bi-LSTM model with tensors and report its detailed evaluation result in the multi-way classification setting. As another baseline, we report the performance of a Bi-LSTM model without tensors as well. Both baseline models take two relevant discourse units as the only input.

For additional comparisons, We also report the performance of our proposed models in the binary classification setting.

4.5 Experimental Results

Multi-way Classification: The first section of table 3 shows macro average F1-scores and accuracies of previous works. The second section of table 3 shows the multi-class classification results of our implemented baseline systems. Consistent with results of previous works, neural tensors, when applied to Bi-LSTMs, improved implicit discourse relation prediction performance. However, the performance on the three small classes (Comp, Cont and Temp) remains low.

The third section of table 3 shows the multi-class classification results of our proposed paragraph-level neural network models that capture inter-dependencies among discourse units. The first row shows the performance of a variant of our basic model, where we only identify implicit relations and ignore identifying explicit relations by setting the α in equation (5) to be 0. Compared with the baseline Bi-LSTM model, the only difference is that this model considers paragraph-wide contexts and model inter-dependencies among discourse units when building representation for individual DU. We can see that this model has greatly improved implicit relation classification perfor-

Model	Comp	Cont	Exp	Temp
(Chen et al., 2016)	40.17	54.76	-	31.32
(Liu et al., 2016)	37.91	55.88	69.97	37.17
(Liu and Li, 2016)	36.70	54.48	70.43	38.84
(Qin et al., 2017)	40.87	54.56	72.38	36.20
(Lei et al., 2017)	40.47	55.36	69.50	35.34
(Lan et al., 2017)	40.73	**58.96**	**72.47**	38.50
Paragraph level Discourse Relation Recognition				
Basic System ($\alpha = 1$)	42.68	55.17	68.94	41.03
+ Untie Parameters	**46.79**	57.09	70.41	**45.61**

Table 4: Binary Classification Results on PDTB. We report F1-scores for implicit discourse relations.

	Implicit		Explicit	
Model	Macro	Acc	Macro	Acc
Basic System ($\alpha = 1$)	49.92	59.08	93.05	93.83
+ Untie Parameters	50.47	**59.85**	93.95	94.74
+ the CRF Layer	**51.84**	59.75	**94.17**	**94.82**

Table 5: Multi-class Classification Results of Ensemble Models on PDTB.

mance across all the four relations and improved the macro-average F1-score by over 7 percents. In addition, compared with the baseline Bi-LSTM model with tensor, this model improved implicit relation classification performance across the three small classes, with clear performance gains of around 2 and 8 percents on contingency and temporal relations respectively, and overall improved the macro-average F1-score by 2.2 percents.

The second row shows the performance of our basic paragraph-level model which predicts both implicit and explicit discourse relations in a paragraph. Compared to the variant system (the first row), the basic model further improved the classification performance on the first three implicit relations. Especially on the contingency relation, the classification performance was improved by another 1.42 percents. Moreover, the basic model yields good performance for recognizing explicit discourse relations as well, which is comparable with previous best result (92.05% macro F1-score and 93.09% accuracy as reported in (Pitler et al., 2008)).

After untying parameters in the softmax prediction layer, implicit discourse relation classification performance was improved across all four relations, meanwhile, the explicit discourse relation classification performance was also improved. The CRF layer further improved implicit discourse relation recognition performance on the three small classes. In summary, our full paragraph-level neural network model achieves the best macro-average F1-score of 48.82% in predicting implicit discourse relations, which outperforms previous neural tensor network models (e.g., (Lei et al., 2017)) by more than 2 percents and outperforms the best previous system (Lan et al., 2017) by 1 percent.

Binary Classification: From table 4, we can see that compared against the best previous systems, our paragraph-level model with untied parameters in the prediction layer achieves F1-score improvements of 6 points on Comparison and 7 points on Temporal, which demonstrates that paragraph-wide contexts are important in detecting minority discourse relations. Note that the CRF layer of the model is not suitable for binary classification.

4.6 Ensemble Model

As we explained in section 4.2, we ran our models for 10 times to obtain stable average performance. Then we also created ensemble models by applying majority voting to combine results of ten runs. From table 5, each ensemble model obtains performance improvements compared with single model. The full model achieves performance boosting of (51.84 - 48.82 = 3.02) and (94.17 - 93.21 = 0.96) in macro F1-scores for predicting implicit and explicit discourse relations respectively. Furthermore, the ensemble model achieves the best performance for predicting both implicit

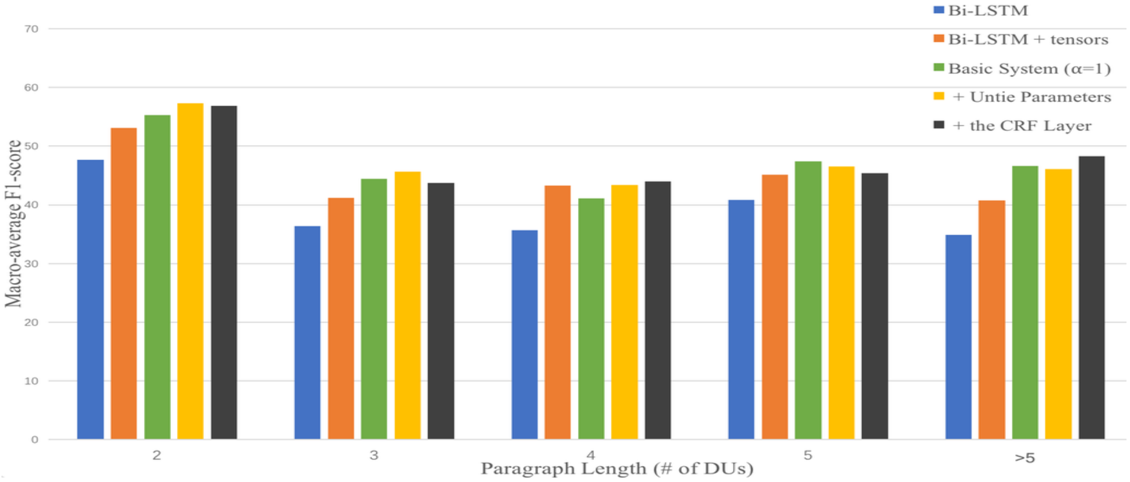

Figure 4: Impact of Paragraph Length. We plot the macro-average F1-score of implicit discourse relation classification on instances with different paragraph length.

and explicit discourse relations simultaneously.

4.7 Impact of Paragraph Length

To understand the influence of paragraph lengths to our paragraph-level models, we divide paragraphs in the PDTB test set into several subsets based on the number of DUs in a paragraph, and then evaluate our proposed models on each subset separately. From Figure 4, we can see that our paragraph-level models (the latter three) overall outperform DU-pair baselines across all the subsets. As expected, the paragraph-level models achieve clear performance gains on long paragraphs (with more than 5 DUs) by extensively modeling mutual influences of DUs in a paragraph. But somewhat surprisingly, the paragraph-level models achieve noticeable performance gains on short paragraphs (with 2 or 3 DUs) as well. We hypothesize that by learning more appropriate discourse-aware DU representations in long paragraphs, our paragraph-level models reduce bias of using DU representations in predicting discourse relations, which benefits discourse relation prediction in short paragraphs as well.

4.8 Example Analysis

For the example (1), the baseline neural tensor model predicted both implicit relations wrongly ("Implicit-Contingency" between DU2 and DU3; "Implicit-Expansion" between DU3 and DU4), while our paragraph-level model predicted all the four discourse relations correctly, which indicates that paragraph-wide contexts play a key role in implicit discourse relation prediction.

For another example:

(2): *[Marshall came clanking in like Marley's ghost dragging those chains of brigades and air wings and links with Arab despots.]*$_{DU1}$ ***(Implicit-Temporal)*** *[He wouldn't leave]*$_{DU2}$ *until* ***(Explicit-Temporal)*** *[Mr. Cheney promised to do whatever the Pentagon systems analysts told him.]*$_{DU3}$

Our basic paragraph-level model wrongly predicted the implicit discourse relation between DU1 and DU2 to be "Implicit-Comparison", without being able to effectively use the succeeding "Explicit-Temporal" relation. On the contrary, the full model corrected this mistake by modeling discourse relation patterns with the CRF layer.

5 Conclusion

We have presented a paragraph-level neural network model that takes a sequence of discourse units as input, models inter-dependencies between discourse units as well as discourse relation continuity and patterns, and predicts a sequence of discourse relations in a paragraph. By building wider-context informed discourse unit representations and capturing the overall discourse structure, the paragraph-level neural network model outperforms the best previous models for implicit discourse relation recognition on the PDTB dataset.

Acknowledgments

We acknowledge the support of NVIDIA Corporation for their donation of one GeForce GTX TITAN X GPU used for this research.

References

Fatemeh Torabi Asr and Vera Demberg. 2012. Implicitness of discourse relations. In *Coling*. pages 2669–2684.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Current and new directions in discourse and dialogue*, Springer, pages 85–112.

Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Implicit discourse relation detection via a deep architecture with gated relevance network. In *ACL 2016*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. pages 681–691. http://aclanthology.info/papers/D17-1071/d17-1071.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* .

Jerry R Hobbs. 1985. *On the coherence and structure of discourse*. CSLI.

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

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *TACL* 3:329–344. https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/536.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse relation language models. In *Proceedings of NAACL-HLT*. pages 332–342.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*. volume 951, pages 282–289.

Man Lan, Jianxiang Wang, Yuanbin Wu, Zheng-Yu Niu, and Haifeng Wang. 2017. Multi-task attention-based neural networks for implicit discourse relationship representation and identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 1310–1319.

Alex Lascarides and Nicholas Asher. 1993. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and philosophy* 16(5):437–493.

Wenqiang Lei, Xuancong Wang, Meichun Liu, Ilija Ilievski, Xiangnan He, and Min-Yen Kan. 2017. Swim: A simple word interaction model for implicit discourse relation recognition. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. pages 4026–4032. https://doi.org/10.24963/ijcai.2017/562.

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

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '09, pages 343–351. http://dl.acm.org/citation.cfm?id=1699510.1699555.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering* 20(2):151–184.

Yang Liu and Mirella Lapata. 2017. Learning contextually informed representations for linear-time discourse parsing. In *EMNLP*.

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 2016, Austin, Texas, USA, November 1-4, 2016*. pages 1224–1233. http://aclweb.org/anthology/D/D16/D16-1130.pdf.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*. pages 2750–2756. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11831.

Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *CoNLL 2016* page 280.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '09, pages 683–691. http://dl.acm.org/citation.cfm?id=1690219.1690241.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Association for Computational Linguistics, pages 13–16.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *In Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008) Short Papers*.

R. Prasad, N. Dinesh, Lee A., E. Miltsakaki, L. Robaldo, Joshi A., and B. Webber. 2008a. The Penn Discourse Treebank 2.0. In *lrec2008*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008b. The penn discourse treebank 2.0. In *In Proceedings of LREC*.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. A stacking gated neural architecture for implicit discourse relation classification. In *EMNLP*. pages 2263–2270.

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric P. Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*. pages 1006–1017. https://doi.org/10.18653/v1/P17-1093.

Attapol Rutherford and Nianwen Xue. 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *HLT-NAACL*. pages 799–808.

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

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

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*. pages 3776–3784.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. pages 3104–3112.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi PrasadO Christopher Bryant, and Attapol T Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. *CoNLL 2015* page 1.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow convolutional neural network for implicit discourse relation recognition. In *EMNLP*. The Association for Computational Linguistics, pages 2230–2235.

Ruqing Zhang, Jiafeng Guo, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2017a. Spherical paragraph model. *arXiv preprint arXiv:1707.05635* .

Yizhe Zhang, Dinghan Shen, Guoyin Wang, Zhe Gan, Ricardo Henao, and Lawrence Carin. 2017b. Deconvolutional paragraph representation learning. In *Advances in Neural Information Processing Systems*. pages 4170–4180.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. volume 2, pages 207–212.