Shallow Convolutional Neural Network for Implicit Discourse Relation Recognition

Biao Zhang[1], **Jinsong Su**[1]*, **Deyi Xiong**[2], **Yaojie Lu**[1], **Hong Duan**[1] and **Junfeng Yao**[1]
Xiamen University, Xiamen, China 361005[1]
Soochow University, Suzhou, China 215006[2]
{zb, lyj}@stu.xmu.edu.cn, {jssu, hduan, yao0010}@xmu.edu.cn
dyxiong@suda.edu.cn

Abstract

Implicit discourse relation recognition remains a serious challenge due to the absence of discourse connectives. In this paper, we propose a Shallow Convolutional Neural Network (SCNN) for implicit discourse relation recognition, which contains only one hidden layer but is effective in relation recognition. The shallow structure alleviates the overfitting problem, while the convolution and nonlinear operations help preserve the recognition and generalization ability of our model. Experiments on the benchmark data set show that our model achieves comparable and even better performance when comparing against current state-of-the-art systems.

1 Introduction

As a crucial task for discourse analysis, discourse relation recognition (DRR) aims to automatically identify the internal structure and logical relationship of coherent text (e.g., TEMPORAL, CONTINGENCY, EXPANSION, etc). It provides important information to many other natural language processing systems, such as question answering (Verberne et al., 2007), information extraction (Cimiano et al., 2005), machine translation (Guzmán et al., 2014) and so on. Despite great progress in explicit DRR where the discourse connectives (e.g., "*because*", "*but*" et al.) explicitly exist in the text (Miltsakaki et al., 2005; Pitler et al., 2008), implicit DRR remains a serious challenge because of the absence of discourse connectives (Prasad et al., 2008).

Conventional methods for implicit DRR directly rely on feature engineering, wherein researchers generally exploit various hand-crafted features, such as words, part-of-speech tags and production rules (Pitler et al., 2009; Lin et al., 2009; Louis et al., 2010; Wang et al., 2012; Park and Cardie, 2012; McKeown and Biran, 2013; Lan et al., 2013; Versley, 2013; Braud and Denis, 2014; Rutherford and Xue, 2014). Although these methods have proven successful, these manual features are labor-intensive and weak to capture intentional, semantic and syntactic aspects that govern discourse coherence (Li et al., 2014), thus limiting the effectiveness of these methods.

Recently, deep learning models have achieved remarkable results in natural language processing (Bengio et al., 2003; Bengio et al., 2006; Socher et al., 2011b; Socher et al., 2011a; Socher et al., 2013; Li et al., 2013; Kim, 2014). However, to the best of our knowledge, there is little deep learning work specifically for implicit DRR. The neglect of this important domain may be due to the following two reasons: (1) discourse relation distribution is rather unbalanced, where the generalization of deep models is relatively insufficient despite their powerful studying ability; (2) training dataset in implicit DRR is relatively small, where overfitting problems become more prominent.

In this paper, we propose a Shallow Convolutional Neural Network (SCNN) for implicit DRR, with only one simple convolution layer on the top of word vectors. On one hand, the network structure is simple, thereby overfitting issue can be alleviated; on the other hand, the convolution operation and nonlinear transformation help preserve the recognition ability of SCNN. This makes our model able to generalize better on the test dataset. We performed evaluation for English implicit DRR on the PDTB-style corpus. Experimental results show that the proposed method can obtain comparable even better performance when compares against several baselines.

2 Model

In Penn Discourse Treebank (PDTB) (Prasad et al., 2008), implicit discourse relations are anno-

*Corresponding author

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2230–2235,
Lisbon, Portugal, 17-21 September 2015.

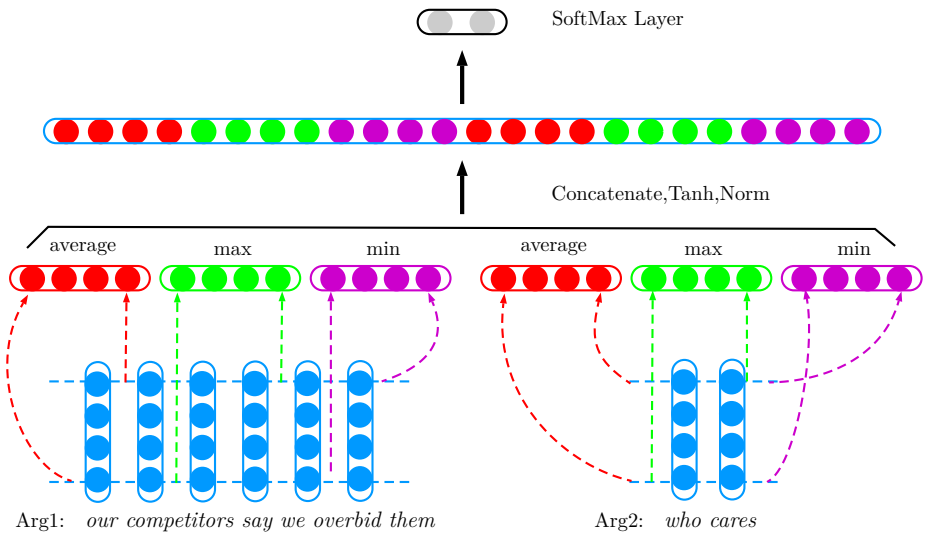

Figure 1: SCNN model architecture visualized with an instance.

tated with connective expressions that best convey implicit relations between two neighboring arguments, e.g.

Arg1: (But) *our competitions say we overbid them*

Arg2: *who cares*

the connective "*But*", which is annotated manually, is used to express the inferred COMPARISON relation.

We learn a classifier for implicit DRR based on convonlutional neural network. The overall model architecture is illustrated in Figure 1.[1] In our model, each word in vocabulary V corresponds to a d-dimensional dense, real-valued vector, and all words are stacked into a word embedding matrix $L \in \mathbb{R}^{d \times |V|}$, where $|V|$ is the vocabulary size.

Given an ordered list of n words in an argument, we retrieve the i-th word representation $x_{v_i} \in \mathbb{R}^d$ from L with its corresponding vocabulary index v_i. All word vectors in the argument produce the following output matrix:

$$\mathrm{X} = (x_{v_1}, x_{v_2}, \ldots, x_{v_n}) \quad (1)$$

Following previous work (Collobert et al., 2011; Socher et al., 2011a), for each row r in X, we explore three convolutional operations to obtain three convolution features *average*, *min* and *max* as follows:

$$c_r^{avg} = \frac{1}{n} \sum_i^n X_{r,i} \quad (2)$$

$$c_r^{min} = \min\left(X_{r,1}, X_{r,2}, \ldots, X_{r,n}\right) \quad (3)$$

$$c_r^{max} = \max\left(X_{r,1}, X_{r,2}, \ldots, X_{r,n}\right) \quad (4)$$

In this way, SCNN is able to capture almost all important information inside X (one with the highest, lowest and average values). Besides, each convolution operation naturally deals with variable argument lengths (Note that $c \in \mathbb{R}^d$). Back to Figure 1, we present c^{avg}, c^{min} and c^{max} with red, purple and green color respectively.

After obtaining the features of both arguments, we concatenate all of them into one vector, and then apply $tanh$ transformation and length normalization successively to generate the hidden layers:

$$a = \left[c_{Arg1}^{avg}; c_{Arg1}^{min}; c_{Arg1}^{max}; c_{Arg2}^{avg}; c_{Arg2}^{min}; c_{Arg2}^{max}\right] \quad (5)$$

$$h = \frac{\tanh(a)}{\|\tanh(a)\|} \quad (6)$$

where $h \in \mathbb{R}^{6d}$ is the hidden layer representation. The normalization operation scales the components of a feature vector to unit length. This, to some extent, eliminates the manifold differences among different features.

Upon the hidden layer, we stack a Softmax layer for relation recognition,

$$y = f(Wh + b) \quad (7)$$

where f is the softmax function, $W \in \mathbb{R}^{l \times 6d}$ is the parameter matrix, $b \in \mathbb{R}^l$ is the bias term, and l is the relation number.

To assess how well the predicted relation y represents the real relation, we supervise it with the

[1] For better illustration, we assume that the dimension of word vectors is 4 throughout this paper.

gold relation g in the annotated training corpus using the traditional cross-entropy error,

$$E(y, g) = -\sum_{j}^{l} g_j \times \log(y_j) \quad (8)$$

Combined with the regularization error, the joint training objective function is

$$J(\theta) = \frac{1}{m}\sum_{t=1}^{m} E(y_t, g_t) + \frac{\lambda}{2}\|\theta\|^2 \quad (9)$$

where m is the number of training instances, y_t is the t-th predicted distribution, λ is the regularization coefficient and θ is parameters, including L, W and b.[2]

To train SCNN, we first employ the toolkit *Word2Vec*[3] (Mikolov et al., 2013) to initialize the word embedding matrix L using a large-scale unlabeled data. Then, L-BFGS algorithm is applied to fine-tune the parameters θ.

3 Experiments

We conducted a series of experiments on English implicit DRR task. After a brief description of the experimental setup and the baseline systems, we further investigated the effectiveness of our method with deep analysis.

3.1 Setup

For comparison with other systems, we formulated the task as four separate one-against-all binary classification problems: one for each top level sense of implicit discourse relations (Pitler et al., 2009).

We used the *PDTB 2.0* corpus[4] (Prasad et al., 2008) for evaluation. The PDTB corpus contains discourse annotations over 2,312 Wall Street Journal articles, and is organized in different sections. Following Pitler et al. (2009), we used sections 2-20 as training set, sections 21-22 as test set, and sections 0-1 as development set for parameter optimization. For each relation, we randomly extracted the same number of positive and negative instances as training data, while all instances in sections 21 and 22 are used as our test set. The statistics of various data sets is listed in Table 1.

We tokenized PDTB corpus using *Stanford NLP Tool*[5]. For all experiments, we empirically set

[2]The bias terms b is not regularized. We preserve it in the equation just for clarity.

[3]https://code.google.com/p/word2vec/

[4]http://www.seas.upenn.edu/ pdtb/

[5]http://nlp.stanford.edu/software/corenlp.shtml

Relation	Positive/Negative Sentences		
	Train	Dev	Test
COMP.	1942/1942	197/986	152/894
CONT.	3342/3342	295/888	279/767
EXP.	7004/7004	671/512	574/472
TEMP.	760/760	64/1119	85/961

Table 1: Statistics of positive and negative instances in training (Train), development (Dev) and test (Test) sets. COMP.=COMPARISON, CONT.=CONTINGENCY, EXP.=EXPANSION and TEMP.=TEMPORAL

d=128 and λ=$1e^{-4}$. Besides, the unlabeled data for word embedding initialization contains 1.02M sentences with 33.5M words.

3.2 Baselines

We compared our model against the following baseline methods:

- **SVM:** This method learns a support vector machine (SVM) classifier with the labeled data.
- **TSVM:** This method learns a transductive SVM (TSVM) classifiers given the labeled data and unlabeled data. We extracted unlabeled data from above-mentioned 1.02M sentences. After filtering the noise ones, we finally obtained 0.11M unlabeled instances, each of which contains only two clauses.
- **RAE:** This method learns a recursive autoencoder (RAE) classifier with the labeled data. We first utilized standard RAEs to represent arguments, and then stacked a Softmax layer upon them. The hyperparameters were set as follows: word dimension 64, balance factor for reconstruction error 0.10282 and regularization factor $1e^{-5}$. Word embeddings are initialized via Word2Vec.

Rutherford and Xue (2014) show that Brown cluster pair feature is very impactful in implicit DRR (Rutherford and Xue, 2014). This feature is superior to one-hot representation for the interactions between two arguments, such as cross-argument word pair features in our baseline methods. We therefore conducted two additional experiments for comparison:

- **Add-Bro:** This method learns an SVM classifier using baseline system features along with the Brown cluster pair feature.
- **No-Cro:** This method learns an SVM classifier on Add-Bro's features without cross-

Relation	Model	Precision	Recall	Accuracy	MacroF1
COMP. vs Other	**SVM**	22.22	60.53	63.48	32.51
	TSVM	20.53	66.45	57.74	31.37
	Add-Bro	22.79	64.47	63.10	33.68
	No-Cro	22.89	67.76	62.14	**34.22**
	RAE	18.38	62.50	54.21	28.40
	SCNN-No-Norm	21.07	54.61	63.67	30.40
	SCNN	22.00	67.76	60.42	33.22
CONT. vs Other	**SVM**	39.70	67.03	64.05	49.87
	TSVM	38.72	67.03	62.91	49.08
	Add-Bro	39.14	72.40	62.62	50.82
	No-Cro	39.50	74.19	62.81	51.56
	RAE	37.55	68.10	61.28	48.41
	SCNN-No-Norm	39.02	71.33	62.62	50.44
	SCNN	39.80	75.29	63.00	**52.04**
EXP. vs Other	**SVM**	66.35	60.10	61.38	63.07
	TSVM	66.48	61.15	61.76	63.70
	Add-Bro	65.89	58.89	60.71	62.19
	No-Cro	66.73	61.15	61.95	63.82
	RAE	58.24	70.29	56.02	63.67
	SCNN-No-Norm	59.39	74.39	58.03	66.05
	SCNN	56.29	91.11	56.30	**69.59**
TEMP. vs Other	**SVM**	15.76	68.24	67.78	25.61
	TSVM	16.26	77.65	65.68	26.88
	Add-Bro	15.10	68.24	66.25	24.73
	No-Cro	13.89	64.71	64.53	22.87
	RAE	10.02	60.00	52.96	17.17
	SCNN-No-Norm	18.26	67.06	72.94	28.71
	SCNN	20.22	62.35	76.95	**30.54**

Table 2: Performance comparison of different systems on the test set.

argument word pair features.

In addition, to further verify the effectiveness of normalization, we also compared against SCNN model without normalization (**SCNN-No-Norm**).

Throughout our experiments, we used the toolkit *SVM-light*[6] (Joachims, 1999) in all the SVM-related experiments. Following previous work (Pitler et al., 2009; Lin et al., 2009), we adopted the following features for baseline methods:

Bag of Words: Three binary features that check whether a word occurs in *A*rg1, *A*rg2 and both arguments.

Cross-Argument Word Pairs: We group all words from *A*rg1 and *A*rg2 into two sets W_1,W_2 respectively, then extract any possible word pair $(w_i, w_j)(w_i \in W_1, w_j \in W_2)$ as features.

Polarity: The count of positive, negated positive, negative and neutral words in *A*rg1 and *A*rg2 according to the MPQA corpus (English). Their cross products are used as features.

First-Last, First3: The first and last words of each argument, the pair of the first words in two arguments, the pair of the last words in two arguments, and the first three words of each argument are used as features.

Production Rules: We extract all production rules from syntactic trees of arguments. We defined three binary features for each rule to check whether this rule appear in *A*rg1, *A*rg2 and both arguments.

Dependency Rules: We also extracted all dependency rules from dependency trees of arguments. Similarly, we defined three binary features for each rule to check whether this rule appear in *A*rg1, *A*rg2 and both arguments.

In order to collect bag of words, production rules, dependency rules, and cross-argument word pairs, we used a frequency cutoff of 5 to remove rare features, following Lin et al. (2009).

3.3 Results and Analysis

All models are evaluated by assessing the accuracy and F1 scores on account of the imbalance in test set. Besides, for better analysis, we also provided the precision and recall results.

Table 2 summarizes the performance of different models. On the whole, the F1 scores for implicit DRR are relatively low on average: COMP., CONT., EXP. and TEMP. about 32%, 50%, 65% and 28% respectively. This illustrates the difficulty in implicit DRR. Although we ex-

[6]http://svmlight.joachims.org/

pected unlabeled data could obtain improvement, we observed negative results appeared in **TSVM**: COMP. and CONT. dropped 1.14% and 0.79% respectively[7]. The F1 scores of TEMP. and EXP. are improved (1.27% and 0.63% respectively). The main reason may be that our unlabeled data is not strictly from the discourse corpus.

Incorporating Brown cluster pair features enhances the recognition of COMP. and CONT.. Particllarly, **No-Cro** achieves the best result in COMP. 34.22%. But we found no consistent improvement in EXP. and TEMP.: **No-Cro** loses 2.74% in TEMP.; **Add-Bro** loses 0.88% and 2.12% in EXP. and TEMP. respectively. This result is inconsistent with the finding of Rutherford and Xue (2014). The reason may lie in the training strategy, where we used sampling to solve the problem of unbalanced dataset while they reweighted training samples.

Compared with SVM-based models, **RAE** performs poorly in three relations, except EXP. which has the largest training dataset. Maybe **RAE** needs more labeled training data for better results. However, SCNN models perform remarkably well, producing comparable and even better results. Without normalization, **SCNN-No-Norm** gains 0.57%, 2.98% and 3.1% F1 scores for CONT., EXP. and TEMP. respectively, but loses 2.11% for COMP.. We obtain further improvement using **SCNN** with normalization: 0.71%, 2.17%, 6.52% and 4.93% for COMP., CONT., EXP. and TEMP. respectively. This suggests that normalization is useful for generalization of shallow models.

From Table 2, we found that our models do not achieve consistent improvements in precision, but benefit greatly from the gains of recall. Besides, our model works quite well for small dataset (Both accuracy and F1 score are improved in TEMP.). All of these demonstrate that our model is suitable for implicit DRR.

4 Conclusion and Future Work

In this paper, we have presented a convolutional neural network based approach to learn better DRR classifiers. The method is simple but effective for relation recognition. Experiment results show that our approach achieves satisfactory performance against the baseline models.

In the future, we will verify our model on other languages, for example, Chinese and Arabic. Besides, since our model is general to classification problems, we would like to investigate its effectiveness on other similar tasks, such as sentiment classification and movie review classification, etc.

[7]Without special illustration, all improvements and declines are against **SVM**.

Acknowledgments

The authors were supported by National Natural Science Foundation of China (Grant Nos 61303082 and 61403269), Natural Science Foundation of Jiangsu Province (Grant No. BK20140355), Natural Science Foundation of Fujian Province of China (Grant No. 2013J01250), the Special and Major Subject Project of the Industrial Science and Technology in Fujian Province 2013 (Grant No. 2013HZ0004-1), and 2014 Key Project of Anhui Science and Technology Bureau (Grant No. 1301021018). We thank the anonymous reviewers for their insightful comments. We are also grateful to Kaixu Zhang for his valuable suggestions.

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *JMLR*, 3:1137–1155.

Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. 2006. Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer Berlin Heidelberg.

Chloé Braud and Pascal Denis. 2014. Combining natural and artificial examples to improve implicit discourse relation identification. In *Proc. of COLING*, pages 1694–1705.

Philipp Cimiano, Uwe Reyle, and Jasmin Šarić. 2005. Ontology-driven discourse analysis for information extraction. *Data & Knowledge Engineering*, pages 59–83.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *JMLR*, pages 2493–2537.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proc. of ACL*, pages 687–698. Association for Computational Linguistics.

T. Joachims. 1999. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proc. of EMNLP*, pages 1746–1751. Association for Computational Linguistics.

Man Lan, Yu Xu, and Zhengyu Niu. 2013. Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition. In *Proc. of ACL*, pages 476–485. Association for Computational Linguistics.

Peng Li, Yang Liu, and Maosong Sun. 2013. Recursive autoencoders for ITG-based translation. In *Proc. of EMNLP*, pages 567–577. Association for Computational Linguistics.

Jiwei Li, Rumeng Li, and Eduard Hovy. 2014. Recursive deep models for discourse parsing. In *Proc. of EMNLP*, pages 2061–2069. Association for Computational Linguistics.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proc. of EMNLP*, pages 343–351. Association for Computational Linguistics.

Annie Louis, Aravind Joshi, Rashmi Prasad, and Ani Nenkova. 2010. Using entity features to classify implicit discourse relations. In *Proc. of SIGDIAL*, pages 59–62. Association for Computational Linguistics.

Kathleen McKeown and Or Biran. 2013. Aggregated word pair features for implicit discourse relation disambiguation. In *Proc. of ACL*, pages 69–73. The Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546.

Eleni Miltsakaki, Nikhil Dinesh, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2005. Experiments on sense annotations and sense disambiguation of discourse connectives. In *Proc. of TLT2005*.

Joonsuk Park and Claire Cardie. 2012. Improving Implicit Discourse Relation Recognition Through Feature Set Optimization. In *Proc. of SIGDIAL*, pages 108–112. Association for Computational Linguistics.

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

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proc. of ACL-AFNLP*, pages 683–691. Association for Computational Linguistics.

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

Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proc. of EACL*, pages 645–654. Association for Computational Linguistics.

Richard Socher, Eric H. Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y. Ng. 2011a. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Proc. of NIPS*, pages 801–809. Curran Associates, Inc.

Richard Socher, Cliff Chiung-Yu Lin, Andrew Y. Ng, and Christopher D. Manning. 2011b. Parsing natural scenes and natural language with recursive neural networks. In *Proc. of ICML*, pages 129–136. Omnipress.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. of EMNLP*, pages 1631–1642. Association for Computational Linguistics.

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2007. Evaluating discourse-based answer extraction for why-question answering. In *Proc. of SIGIR*, pages 735–736. ACM.

Yannick Versley. 2013. Subgraph-based classification of explicit and implicit discourse relations. In *Proc. of IWCS*, pages 264–275. Association for Computational Linguistics.

Xun Wang, Sujian Li, Jiwei Li, and Wenjie Li. 2012. Implicit discourse relation recognition by selecting typical training examples. In *Proc. of COLING*, pages 2757–2772.