A Stacking Gated Neural Architecture for Implicit Discourse Relation Classification

Lianhui Qin[1,2], **Zhisong Zhang**[1,2], **Hai Zhao**[1,2,*]
[1]Department of Computer Science and Engineering,
Shanghai Jiao Tong University, Shanghai, 200240, China
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
{qinlianhui, zzs2011}@sjtu.edu.cn,zhaohai@cs.sjtu.edu.cn

Abstract

Discourse parsing is considered as one of the most challenging natural language processing (NLP) tasks. Implicit discourse relation classification is the bottleneck for discourse parsing. Without the guide of explicit discourse connectives, the relation of sentence pairs are very hard to be inferred. This paper proposes a stacking neural network model to solve the classification problem in which a convolutional neural network (CNN) is utilized for sentence modeling and a collaborative gated neural network (CGNN) is proposed for feature transformation. Our evaluation and comparisons show that the proposed model outperforms previous state-of-the-art systems.

1 Introduction

As a fundamental task in natural language processing (NLP), discourse parsing entails the discovery of the latent relational structure in multi-sentence level analysis. It is also central to many practical tasks such as question answering (Liakata et al., 2013; Jansen et al., 2014), machine translation (Meyer and Popescu-Belis, 2012; Meyer and Webber, 2013) and automatic summarization (Murray et al., 2006; Yoshida et al., 2014). Discourse parsing is also the shared task of CoNLL 2015 and 2016 (Xue et al., 2015; Xue et al., 2016), and many previous works previous on this task (Qin et al., 2016b; Li et al., 2016; Chen et al., 2015; Wang and Lan, 2016). In a discourse parser, implicit relation recognition has been the bottleneck due to lack of explicit connectives (like *"because"* or *"and"*) that can be strong indicators for the senses between adjacent clauses (Qin et al., 2016b; Pitler et al., 2009; Lin et al., 2014). This work therefore focuses on implicit relation recognition that infers the senses of the discourse relations within adjacent sentence pairs.

Most previous works on PDTB implicit relation recognition only focus on one-versus-others binary classification problems of the top level four classes (Pitler et al., 2009; Zhou et al., 2010; Park and Cardie, 2012; Biran and McKeown, 2013; Rutherford and Xue, 2014; Braud and Denis, 2015). Traditional classification methods directly rely on feature engineering, based on bag-of-words, production rules, and some linguistically-informed features (Zhou et al., 2010; Rutherford and Xue, 2014). However, discourse relations root in semantics, which may be hard to recover from surface level feature, thus these methods did not report satisfactory performance. Recently, neural network (NN) models have shown competitive or even better results than traditional linear models with handcrafted sparse features (Wang et al., 2016b; Zhang et al., 2016a; Jia and Zhao, 2014). They have been proved to be effective for many tasks (Qin et al., 2016a; Wang et al., 2016a; Zhang et al., 2016b; Wang et al., 2015; Wang et al., 2014; Cai and

*Corresponding author. This paper was partially supported by Cai Yuanpei Program (CSC No. 201304490199 and No. 201304490171), National Natural Science Foundation of China (No. 61170114, No. 61672343 and No. 61272248), National Basic Research Program of China (No. 2013CB329401), Major Basic Research Program of Shanghai Science and Technology Committee (No. 15JC1400103), Art and Science Interdisciplinary Funds of Shanghai Jiao Tong University (No. 14JCRZ04), and Key Project of National Society Science Foundation of China (No. 15-ZDA041).

Zhao, 2016), also including discourse parsing. Ji and Eisenstein (2015) adopt recursive neural network and incorporate with entity-augmented distributed semantics. Zhang et al. (2015) explore a shallow convolutional neural network and achieve competitive performance. Although simple neural network has been shown effective, the result has not been quite satisfactory which suggests that there is still space for improving.

The concerned task could be straightforwardly formalized as a sentence-pair classification problem, which needs inferring senses solely based on the two arguments without cues of connectives. Two problems should be carefully handled in this task: how to model sentences and how to capture the interactions between the two arguments. The former could be addressed by Convolutional Neural Network (CNN) which has been proved effective for sentence modeling (Kalchbrenner et al., 2014; Kim, 2014), while the latter is the key problem, which might need deep semantic analysis for the interaction of two arguments. To solve the latter problem, we propose collaborative gated neural network (CGNN) which is partially inspired by Highway Network whose gate mechanism achieves success (Srivastava et al., 2015). Our method will be evaluated on the benchmark dataset against state-of-the-art methods.

The rest of the paper is organized as follows: Section 2 briefly describes our model, introducing the stacking architecture of CNN and CGNN, Section 3 shows the experiments and analysis, and Section 4 concludes this paper.

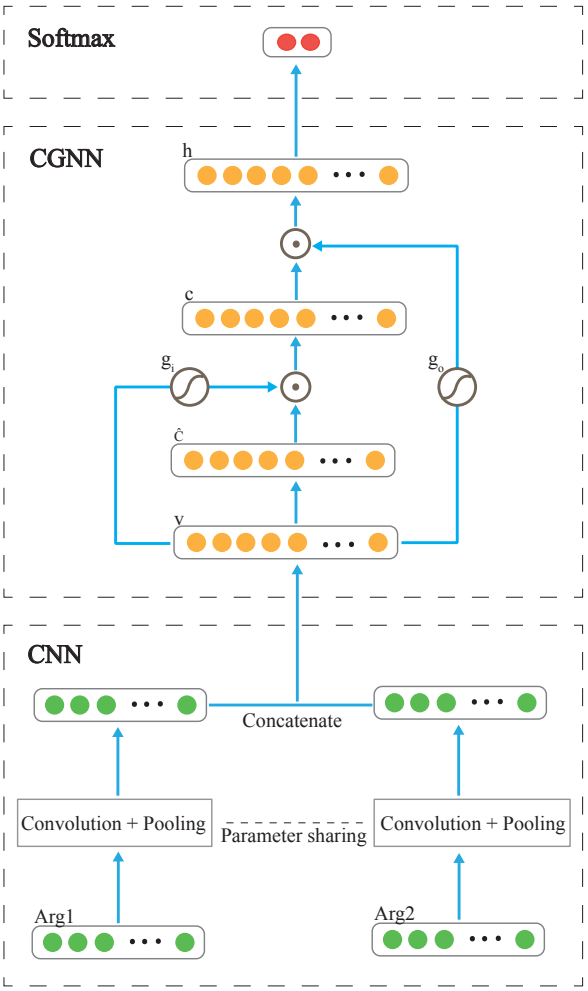

Figure 1: Model Architecture

2 Method

The architecture of the model, as shown in Figure 1, is straightforward. It can be divided into three parts: 1) CNN for modeling arguments; 2) CGNN unit for feature transformation; 3) a conventional softmax layer for the final classification. CNN is used to obtain the vector representations for the sentences, CGNN further captures and transforms the features for the final classification.

2.1 Convolutional Neural Network

As CNN has been broadly adopted for modeling sentences, we will explain it in brevity. For two arguments, typical sentence modeling process will be applied: sentence embedding (including embeddings for words and part-of-speech (POS) tags) through projection layer, convolution operations (with multiple groups of filters) through the convolution layer, obtaining the sentence representation through one-max-pooling. The two arguments will get their sentence vectors independently without any interfering, and the convolution operation will be the same by sharing parameters. The final argument-pair representation will be the vector $\mathbf{v}$ which is concatenated from two sentence vectors and this vector will be used as the input of the CGNN unit.

	COMP.	CONT.	EXP.	TEMP.	AVG
CNN Only	39.07	54.73	65.94	30.19	47.48
CNN+MLP	37.81	56.30	69.44	32.29	48.96
CNN+LSTM	39.15	53.44	68.85	29.79	47.81
CNN+Highway	37.72	56.35	68.94	30.56	48.39
CNN+CGNN	**41.55**	**57.32**	**71.50**	**35.43**	**51.45**

Table 1: F_1 scores (%) with different models.

2.2 Collaborative Gated Neural Network

For implicit sense classification, the key is how to effectively capture the interactions between the two arguments. The interactions could be word pairs, phrase pairs or even the latent meaning of the two full arguments. Pitler et al. (2009) has shown that word pair features are helpful. To model these interactions, we have to make a full use of the sentence vectors obtained from CNN. However, common neural hidden layers might be insufficient to deal with the challenge. We need to seek more powerful neural models, i.e., gated neural network.

In recent years, gated mechanism has gained popularity in neural models. Although it is first introduced in the cells of recurrent neural networks, like Long-Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Unit (GRU) (Chung et al., 2014), traditional feed-forward neural models such as the Highway Network could also benefit from it (Srivastava et al., 2015). The existing studies show that the gated mechanism in highway network serves not only a means for easier training, but also a tool to route information in a trained network.

Motivated by the idea of highway network, we propose a collaborative gated neural network (CGNN) for this task. The architecture of CGNN is illustrated in Figure 1, and it contains a sequence of transformations. First, the inner-cell $\hat{\mathbf{c}}$ is obtained through linear transformation and non-linear activation on the input $\mathbf{v}$, and this process is exactly the operation of an ordinary neural layer.

$$\hat{\mathbf{c}} = \tanh(\mathbf{W}^c \cdot \mathbf{v} + \mathbf{b}^c)$$

Meanwhile, the two gates $\mathbf{g}_i$ and $\mathbf{g}_o$ are calculated independently because they are only influenced by the original input through different parameters:

$$\mathbf{g}_i = \sigma(\mathbf{W}^i \cdot \mathbf{v} + \mathbf{b}^i)$$
$$\mathbf{g}_o = \sigma(\mathbf{W}^o \cdot \mathbf{v} + \mathbf{b}^o)$$

where the σ denotes sigmoid function which guarantees the values in the gates are in [0,1]. Two gated operations are applied sequentially, where a gated operation indicates the element-wise multiplication of an inner-cell and a gate. Between the two gated operations, a non-linear activation operation is applied. The procedure could be formulated as follows:

$$\mathbf{c} = \hat{\mathbf{c}} \odot \mathbf{g}_i$$
$$\mathbf{h} = \tanh(\mathbf{c}) \odot \mathbf{g}_o$$

where $\odot$ denotes element-wise multiplication, $\mathbf{c}$ is the second inner-cell and $\mathbf{h}$ is the output of CGNN unit.

Although the two gates are generated independently, they will work collaboratively because they control the information flow of the inner-cells sequentially which resembles logical AND operation in a probabilistic version. In fact, the transformations after $\hat{\mathbf{c}}$ will concern only element-wise operations which might give finer controls for each dimension, and the information can only flow on the dimensions where both gates are "open". This procedure will help select the most crucial features.

The gates in this model are mainly used for routing information from sentence-pairs vectors. When there is only one gate in our network, the model works similar to the highway network (Srivastava et al., 2015).

2.3 Output and Training

After the transformation of the CGNN unit, the transformed vector $\mathbf{h}$ will be sent to a conventional softmax for classification.

The training object *J* will be the cross-entropy error *E* with *L2* regularization:

$$E(\hat{y}, y) = -\sum_{j}^{l} y_j \times \log(Pr(\hat{y}_j))$$

$$J(\theta) = \frac{1}{m}\sum_{k}^{m} E(\hat{y}^{(k)}, y^{(k)}) + \frac{\lambda}{2}\|\theta\|^2$$

where y_j is the gold label and $\hat{y}_j$ is the predicted one. We adopt the diagonal variant of AdaGrad (Duchi et al., 2011) for the optimization process.

3 Experiments

3.1 Setting

As for the benchmark dataset, Penn Discourse Treebank (PDTB) (Prasad et al., 2008) corpus[1] is used for evaluation. In the PDTB, each discourse relation is annotated between two argument spans.

To be consistent with the setups of prior works, we formulate the implicit relation classification task as four one-versus-other binary classification problems only using the four top level classes: COMPARISON (COMP.), CONTINGENCY (CONT.), EXPANSION (EXP.) and TEMPORAL (TEMP.). While different works include different relations of varying specificities, all of them include these four core relations (Pitler et al., 2009). Following dataset splitting convention of the previous works, we use sections 2-20 for training, sections 21-22 for testing and sections 0-1 for development set. The proposed model is possible to be extended for multi-class classification of discourse parsing, but for the comparisons with most of previous works, we will follow them and focus on the binary classification problems.

For other hyper-parameters of the model and training process, we fix the lengths of both the input arguments to be 80, and apply truncating or zero-padding when necessary. The dimensions for word embeddings and POS embeddings are respectively 300 and 50, and the embedding layer adopts a dropout of 0.2. The word embeddings are initialized with pre-trained word vectors using *word2vec* [2] (Mikolov et al., 2013) and other parameters are randomly initialized including POS embeddings. We set the starting learning rate to 0.001. For CNN model, we utilize three groups of filters with window widths of (2, 2, 2) and their filter numbers are all set to 1024. The hyper-parameters are the same for all models and we do not tune them individually.

3.2 Model Analysis

For transformation of sentence vectors, a simple Multilayer Perceptron (MLP) layer could be a straightforward choice, while more complex neural modules, such as LSTM and highway network, could also be considered. Our model utilizes a CGNN unit with refined gated mechanism for the transformation. Will the proposed CGNN really bring about further performance improvement? We now answer this question empirically.

As shown in Table 1, CNN model usually performs well on its own. Utilizing an MLP layer or a Highway layer could improve the accuracies on CONTINGENCY, EXPANSION, TEMPORARY except for COMPARISON. Though the primary motivation of Highway is to ease gradient-based training of highly deep networks through utilizing gated units, it works merely as an ordinary MLP in the proposed model, which explains the reason that it performs like MLP. Despite one of four classes, COMPARISON, not receiving performance improvement, introducing a non-linear transformation layer lets the classification benefit as a whole. "CNN+LSTM" denotes the method of using LSTM to read the convolution sequence (without pooling operation), and it even does not perform better than MLP.

The CGNN achieves the best performance on all classes including COMPARISON. It gains 3.97% imrovement on average F1 score using CNN only model. We assume that CGNN is well-suited to work with CNN, adaptively transforming and combining local features detected by the individual filters.

3.3 Results

We show the main results in Tables 2 and 3. The metrics include precision (P), recall (R), accuracy (Acc) and F1 score. Since not all of these metrics are reported in previous work, the comparisons are correspondingly in Table 2 and 3. Some previous work merges *Entrel* with *Expansion*, which is also explored in our study and noted as EXP.+.

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

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

	COMP.		CONT.		EXP.+		TEMP.		AVG.	
	F_1	Acc	F_1	Acc	F_1	Acc	F_1	Acc	F1	Acc
Pitler et al. (2009)	21.96	56.59	47.13	67.30	76.42	63.62	16.76	63.49	40.57	62.75
Zhou et al. (2010)	31.79	58.22	47.16	48.96	70.11	54.54	20.30	55.48	40.32	54.30
P&C (2012)	31.32	74.66	49.82	72.09	79.22	69.14	26.57	79.32	46.73	73.80
M&B (2013)	25.40	63.36	46.94	68.09	75.87	62.84	20.23	68.35	42.11	65.66
J& (2015)	35.93	70.27	52.78	76.95	80.02	69.80	27.63	87.11	49.09	76.03
B&D(2015)	36.36	-	55.76	-	61.76	-	29.30	-	45.80	-
Chen et al. (2016)	40.17	-	54.76	-	80.62	-	31.32	-	51.72	-
Current	**41.55**	71.22	**57.32**	73.80	**80.96**	68.44	**35.43**	84.32	**53.82**	74.45

Table 2: Comparisons of F_1 scores (%) (symbol **+** means EXP. with *Entrel*).

		P	R	F_1
COMP.	R&Xue (2014)	27.34	72.41	39.70
	Zhang et al.(2015)	22.00	67.76	33.22
	Current	29.48	70.39	**41.55**
CONT.	R&Xue (2014)	44.52	69.96	54.42
	Zhang et al.(2015)	39.80	75.29	52.04
	Current	50.69	65.95	**57.32**
EXP.	R&Xue (2014)	59.59	85.50	70.23
	Zhang et al.(2015)	56.29	91.11	69.59
	Current	60.81	86.76	**71.50**
TEMP.	R&Xue (2014)	18.52	63.64	28.69
	Zhang et al.(2015)	20.22	62.35	30.54
	Current	26.63	52.94	**35.43**
AVG.	R&Xue (2014)	37.49	72.88	48.26
	Zhang et al.(2015)	34.58	74.13	46.35
	Current	41.90	69.01	**51.45**

Table 3: Comparisons of F_1 scores (%) (EXP. without *Entrel*).

We compare with best-performed or competitive models including both traditional linear methods and recent neural methods. For traditional methods: Pitler et al. (2009) use several linguistically informed features, including polarity tags, Levin verb classes, length of verb phrases, modality, context, and lexical features; Zhou et al. (2010) improve the performance through predicting connective words as features; Park and Cardie (2012) propose a locally-optimal feature set and further identify factors for feature extraction that can have a major impact performance, including stemming and lexicon look-up; Biran and McKeown (2013) collect word pairs from arguments of explicit examples to help the learning; Rutherford and Xue (2014) employ Brown cluster pair and coreference patterns for performance enhancement. Several neural methods have also been included for comparison: Zhang et al. (2015) propose a simplified neural network which has only three different pooling operations (max, min, average); Ji and Eisenstein (2015) compute distributed semantics representation by composition up the syntactic parse tree through recursive neural network; Braud and Denis (2015) consider shallow lexical features and word embeddings. Chen et al. (2016) replace the original words by word embeddings to overcome the data sparsity problem and they also utilize gated relevance network to capture the semantic interaction between word pairs. The gated network is different from ours but also works well.

Our model achieves F-measure improvements of 1.85% on COMPARISON, 1.56% on CONTINGENCY, 1.27% on EXPANSION, 0.94% on EXPANSION+, 4.89% on TEMPORAL, against the state-of-the-art of each class. We improve by 4.73% on average F1 score when not including ENTREL in EXPANSION as reported in Table 2 and 3.19% on average F1 score otherwise as reported in Table 3. The results show that our model achieves the best performance and especially makes the most remarkable progress on TEMPORAL.

4 Conclusion

In this paper, we propose a stacking gated neural architecture for implicit discourse relation classification. Our model includes convolution and collaborative gated neural network. The analysis and experiments show that CNN performs well on its own and combining CGNN provides further gains. Our evaluation on PTDB shows that the proposed model outperforms previous state-of-the-art systems.

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)*, pages 69–73, Sofia, Bulgaria, August.

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

Deng Cai and Hai Zhao. 2016. Neural Word Segmentation Learning for Chinese. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 409–420, Berlin, Germany, August.

Changge Chen, Peilu Wang, and Hai Zhao. 2015. Shallow discourse parsing using constituent parsing tree. In *Proceedings of the CoNLL-15 shared task*, pages 37–41, Beijing, China, July.

Jifan Chen, Qi Zhang, Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Implicit discourse relation detection via a deep architecture with gated relevance network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1726–1735, Berlin, Germany, August.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural computation*, 9(8):1735–1780.

Peter Jansen, Mihai Surdeanu, and Peter Clark. 2014. Discourse complements lexical semantics for non-factoid answer reranking. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 977–986, Baltimore, Maryland, June.

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.

Zhongye Jia and Hai Zhao. 2014. A Joint Graph Model for Pinyin-to-Chinese Conversion with Typo Correction. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1512–1523, Baltimore, Maryland, June.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 655–665, Baltimore, Maryland, June.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October.

Zhongyi Li, Hai Zhao, Chenxi Pang, Lili Wang, and Huan Wang. 2016. A constituent syntactic parse tree based discourse parser. In *Proceedings of the CoNLL-16 shared task*, pages 60–64, Berlin, Germany, August.

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)*, pages 747–757, Seattle, Washington, USA, October.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A pdtb-styled end-to-end discourse parser. *Natural Language Engineering*, 20(02):151–184.

Thomas Meyer and Andrei Popescu-Belis. 2012. Using sense-labeled discourse connectives for statistical machine translation. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 129–138, Avignon, France, April.

Thomas Meyer and Bonnie Webber. 2013. Implicitation of discourse connectives in (machine) translation. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 19–26, Sofia, Bulgaria, August.

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 (NIPS)*, pages 3111–3119, South Lake Tahoe, Nevada, US, December.

Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore. 2006. Incorporating speaker and discourse features into speech summarization. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (NAACL)*, pages 367–374, New York City, USA, June.

Joonsuk Park and Claire Cardie. 2012. Improving implicit discourse relation recognition through feature set optimization. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 108–112, Seoul, South Korea, July.

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)*, pages 683–691, Suntec, Singapore, August.

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 *Proceedings of the Sixth conference on International Language Resources and Evaluation (LREC-2008)*, pages 2961–2968, Marrakech, Morocco, May.

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

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

Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 645–654, Gothenburg, Sweden, April.

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

Jianxiang Wang and Man Lan. 2016. Two End-to-end Shallow Discourse Parsers for English and Chinese in CoNLL-2016 Shared Task. In *Proceedings of the CoNLL-16 shared task*, pages 33–40, Berlin, Germany, August.

Rui Wang, Hai Zhao, Bao-Liang Lu, Masao Utiyama, and Eiichiro Sumita. 2014. Neural network based bilingual language model growing for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 189–195, Doha, Qatar, October.

Peilu Wang, Yao Qian, Frank K Soong, Lei He, and Hai Zhao. 2015. Word embedding for recurrent neural network based TTS synthesis. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883, Brisbane, Australia.

Peilu Wang, Yao Qian, Frank K. Soong, Lei He, and Hai Zhao. 2016a. 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)*, pages 527–533, San Diego, California, June.

Rui Wang, Masao Utiyama, Isao Goto, Eiichiro Sumita, Hai Zhao, and Bao-Liang Lu. 2016b. Converting continuous-space language models into n-gram language models with efficient bilingual pruning for statistical machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 15(3):11.

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 CoNLL-15 shared task*, pages 1–16, Beijing, China, July.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. CoNLL 2016 Shared Task on Multilingual Shallow Discourse Parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany, August.

Yasuhisa Yoshida, Jun Suzuki, Tsutomu Hirao, and Masaaki Nagata. 2014. Dependency-based discourse parser for single-document summarization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1834–1839, Doha, Qatar, October.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow convolutional neural network for implicit discourse relation recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2230–2235, Lisbon, Portugal, September.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Hai Zhao, Graham Neubig, and Satoshi Nakamura. 2016a. Learning local word reorderings for hierarchical phrase-based statistical machine translation. *Machine Translation*, pages 1–18.

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)*, pages 1382–1392, Berlin, Germany, August.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting discourse con-

nectives for implicit discourse relation recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 1507–1514, Beijing, China, August.