Employing the Correspondence of Relations and Connectives to Identify Implicit Discourse Relations via Label Embeddings

Linh The Nguyen[†], **Linh Van Ngo**[†], **Khoat Than**[†] **and Thien Huu Nguyen**[‡*]
[†] Hanoi University of Science and Technology, Hanoi, Vietnam
[‡] Department of Computer and Information Science, University of Oregon, USA
{linh.nt142585,linhnv,khoattq}@sis.hust.edu.vn,thien@cs.uoregon.edu

Abstract

It has been shown that implicit connectives can be exploited to improve the performance of the models for implicit discourse relation recognition (IDRR). An important property of the implicit connectives is that they can be accurately mapped into the discourse relations conveying their functions. In this work, we explore this property in a multi-task learning framework for IDRR in which the relations and the connectives are simultaneously predicted, and the mapping is leveraged to transfer knowledge between the two prediction tasks via the embeddings of relations and connectives. We propose several techniques to enable such knowledge transfer that yield the state-of-the-art performance for IDRR on several settings of the benchmark dataset (i.e., the Penn Discourse Treebank dataset).

1 Introduction

Discourse parsing reveals the discourse units (i.e., text spans, sentences, clauses) of the documents and how such units are related to each others to improve the coherence. This work focuses on the task of implicit discourse relation recognition (IDRR), aiming to identify the discourse relations (i.e., *cause*, *contrast*) between adjacent text spans in documents. IDRR is a fundamental problem in discourse analysis (Knott, 2014; Webber et al., 1999) with important applications on question answering (Liakata et al., 2013; Jansen et al., 2014) and text summarization (Gerani et al., 2014; Yoshida et al., 2014), to name a few. Due it its importance, IDRR is being studied actively in the literature, leading to the recent advances for this problem based on deep learning (Chen et al., 2016; Qin et al., 2016; Zhang et al., 2016; Lan et al., 2017; Dai and Huang, 2018).

Consider the two following text spans (called arguments) taken from (Qin et al., 2017) as an example:

Argument 1: *Never mind.*

Argument 2: *You already know the answer.*

An IDRR model should be able to recognize that argument 2 is the cause of argument 1 (i.e., the *Cause* relation) in this case. This is a challenging problem as the models need to rely solely on the text of the arguments to predict accurate discourse relations. The problem would become more manageable if connective/marker cues (i.e., "*but*", "*so*") are provided to connect the two arguments according to their discourse relations (Qin et al., 2017). In the example above, it is beneficial for the models to know that "*because*" can be a connective of the two arguments that is consistent with their discourse relation (i.e., *Cause*). In fact, a human annotator can also benefit from the connectives between arguments when he or she needs to assign discourse relations for pairs of arguments (Qin et al., 2017). This is demonstrated in the Penn Discourse Treebank dataset (PDTB) (Prasad et al., 2008), a major benchmark dataset for IDRR, where the annotators first inject the connectives between the arguments (called the "implicit connectives") to aid the relation assignment of the arguments later (Qin et al., 2017).

Motivated by the relevance of connectives for IDRR, some recent work on deep learning has explored methods to transfer the knowledge from the implicit connectives to support discourse relation prediction using the multi-task learning frameworks (Qin et al., 2017; Bai and Zhao, 2018). The typical approach is to simultaneously predict the discourse relations and the implicit connectives for the input arguments in which the model parameters for the two prediction tasks are shared/tied to allow the knowledge transfer (Liu et al., 2016; Wu et al., 2016; Lan et al., 2017;

[*] Corresponding author.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4201–4207
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

Bai and Zhao, 2018). Unfortunately, such multi-task learning models for IDRR share the limitation of failing to exploit the mapping between the implicit connectives and the discourse relations. In particular, each implicit connective in the PDTB dataset can be naturally mapped into the corresponding discourse relations based on their semantics that can be further employed to transfer the knowledge from the connectives to the relations. For instance, in the PDTB dataset, the connective "*consequently*" uniquely corresponds to the relation *cause* while the connective "*in contrast*" can be associated with the relation *comparison*. In this work, we argue that the knowledge transfer facilitated by such a connective-relation mapping can indeed help to improve the performance of the multi-task learning models for IDRR with deep learning. Consequently, in order to exploit the connective-relation mapping, we propose to embed the implicit connectives and the discourse relations into the same space that would be used to transfer the knowledge between connective and relation predictions via the mapping. We introduce several mechanisms to encourage both knowledge sharing and representation distinction for the embeddings of the connectives and relations for IDRR. In the experiments, we extensively demonstrate that the novel embeddings of connectives and relations along with the proposed mechanisms significantly improve the multi-task learning models for IDRR. We achieve the state-of-the-art performance for IDRR over several settings of the benchmark dataset PDTB.

2 Related Work

There have been many research on IDRR since the creation of the PDTB dataset (Prasad et al., 2008). The early work has manually designed various features for IDRR (Pitler et al., 2009; Lin et al., 2009; Wang et al., 2010; Zhou et al., 2010; Braud and Denis, 2015; Lei et al., 2018) while the recent approach has applied deep learning to significantly improve the performance of IDRR (Zhang et al., 2015; Ji et al., 2015a; Chen et al., 2016; Liu et al., 2016; Qin et al., 2016; Zhang et al., 2016; Cai and Zhao, 2017; Lan et al., 2017; Wu et al., 2017; Dai and Huang, 2018; Kishimoto et al., 2018).

The most related work to ours in this paper involves the multi-task learning models for IDRR that employ connectives as the auxiliary labels for the prediction of the discourse relations. For the feature-based approach, (Zhou et al., 2010) employ a pipelined approach to first predict the connectives and then assign discourse relations accordingly while (Lan et al., 2013) use the connective-relation mapping to automatically generate synthetic data. For the recent work on deep learning for IDRR, (Liu et al., 2016; Wu et al., 2016; Lan et al., 2017; Bai and Zhao, 2018) simultaneously predict connectives and relations assuming the shared parameters of the deep learning models while (Qin et al., 2017) develop adversarial networks to encourage the relation models to mimic the features learned from the connective incorporation. However, none of these work employs embeddings of connectives and relations to transfer knowledge with the connective-relation mapping and deep learning as we do in this work.

3 Model

Let A_1 and A_2 be the two input arguments (essentially text spans with sequences of words). The goal of IDRR is to predict the discourse relation r for these two arguments among the n possibilities in the discourse relation set R ($|R| = n$). Following the prior work on IDRR (Qin et al., 2017; Bai and Zhao, 2018), we focus on the PDTB dataset in this work. In PDTB, besides the discourse relation r, each argument pair is also associated with an implicit connective c to aid the relation prediction. The set of possible implicit connectives is denoted as C ($|C| = k$) in PDTB.

In the following, we first describe the multi-task learning framework for IDRR to employ both training signals r and c for A_1 and A_2, and present the novel mechanisms for knowledge transfer with connective and relation embeddings afterward.

3.1 Multi-task Learning for IDRR

The multi-task learning model for IDRR aims to predict the discourse relations and the implicit connectives simultaneously in a single training process so the knowledge from the connective prediction can be transferred to the relation prediction to improve the performance. In particular, the arguments A_1 and A_2 are first consumed by a neural network model M (called the encoder model) to generate a representation vector $V = M(A_1, A_2)$. In the previous work on multi-task learning for IDRR, this representation vector V would be used to compute the probability distributions for both connective and relation predictions based on two task-specific neural networks with softmax layers.

In our multi-task learning model, the representation vector V is also fed into two feed-forward neural networks F_r and F_c to compute the representation vectors V_r and V_c specific to the predictions of the relation r and the connective c respectively (i.e., $V_r = F_r(V) \in \mathbb{R}^d$ and $V_c = F_c(V) \in \mathbb{R}^d$ where d is the dimension of the vectors). However, instead of directly normalizing V_r and V_c with softmax layers, we employ two embedding matrices $E_r \in \mathbb{R}^{n \times d}$ and $E_c \in \mathbb{R}^{k \times d}$ for the relations and connectives respectively. These vectors are multiplied with the representation vectors V_r and V_c to produce the scores for the possibilities, eventually being normalized by the softmax layers to obtain the probability distributions P_r and P_c over the relation set R and the connective set C for prediction: $P_r = \mathit{softmax}(E_r V_r) \in \mathbb{R}^n$ and $P_c = \mathit{softmax}(E_c V_c) \in \mathbb{R}^k$. To train the model, we jointly minimize the negative log-likelihood for the relation r and the connective c:

$$L = -\log(P_r[r]) - \log(P_c[c]) \qquad (1)$$

Note that the embedding matrices E_r and E_c are initialized randomly and updated as model parameters in the training process, and the implicit connectives are only required in the training phase.

The description of the multi-task learning framework so far is agnostic to the encoder model M to generate the vector representation V for A_1 and A_2. In order to ensure a fair comparison with the recent work on multi-task learning for IDRR, in this work, we employ the best encoder model M presented in (Bai and Zhao, 2018), a recent state-of-the-art multi-task learning model for this problem. We refer the reader to (Bai and Zhao, 2018) for the full description of the encoder. Essentially, this encoder first converts the words in the arguments A_1 and A_2 into vectors using the word embedding *word2vec* in (Mikolov et al., 2013b), the word embedding ELMo in (Peters et al., 2018) and the subword embeddings. This transforms the arguments into matrices that are sent to stacks of convolutional neural networks (Nguyen and Grishman, 2015a,b) augmented with gated linear units and residual connections. Each CNN layer produces two hidden matrices corresponding to the two input arguments over which the co-attention and max-pooling mechanisms are applied to obtain a part of the representation vector V with the current CNN layer.

3.2 Knowledge Transferring via Relation and Connective Embeddings

As we have mentioned in the introduction, each implicit connective in C can be associated with a set of discourse relations that capture its main discourse functions. For instance, in the PDTB dataset, we find that 53% of the implicit connectives only corresponds to one discourse relation while the other 44% appears with two discourse relations. Our intuition is to employ such correspondence between connectives and relations to link the similar concepts in the two prediction tasks to promote knowledge transfer. As the connectives and relations are embedded via E_r and E_c in this work, we can rely on such embeddings to enforce the similarity of the corresponding connectives and relations in the training process.

Formally, for each connective $c_i \in C$, let R_i be the relation subset of R that can be paired with c_i in the correspondence. In order to transfer the knowledge from the connective prediction to the relation prediction, we propose to encourage the embedding of c_i to be similar to the embeddings of the relations in R_i, leading to the following loss function to be minimized:

$$L_1 = \sum_{i=1}^{k} \sum_{r_j \in R_i} \|E_c[c_i] - E_r[r_j]\|^2 \qquad (2)$$

where $\|.\|$ is the L_2 norm of a vector, and $E_c[c_i]$ and $E_r[r_j]$ denote the embeddings of the connective c_i and the relation r_j respectively.

The constraint in Equation 2 can have degenerate solutions where the embeddings of the connectives corresponding to some relation all have the same embeddings as the relation embedding. In order to avoid this trivial solution, we propose to add another constraint to ensure that the embeddings of the connectives of the same relation to be different. Formally, for each relation $r_i \in R$, let C_i be the subset of connectives of C that can correspond to r_i and $E_c^{C_i}$ be the matrix containing the embeddings of the connectives in C_i from E_c. We achieve the difference between the embeddings of the connectives by minimizing:

$$L_2 = \sum_{i=1}^{n} \|E_c^{C_i}(E_c^{C_i})^T - I\|_F^2 \qquad (3)$$

where $\|.\|_F$ is the Frobenius norm, I is an identity matrix, and $(E_c^{C_i})^T$ is the transpose matrix of $E_c^{C_i}$.

The objective of the terms in Equation 3 is to encourage the matrices $E_c^{C_i}$ to be orthogonal where each connective embedding captures different semantic aspects (Lin et al., 2017).

Finally, as the discourse relations in IDRR tend to characterize different functions in documents, we apply a similar constraint as Equation 3 on the relation embedding matrix E_r to promote the diversity of the relation embeddings with the following penalization term:

$$L_3 = \|E_r E_r^T - I\|_F^2 \tag{4}$$

Note that the embedding vectors in the matrices $E_c^{C_i}$ and E_r in Equations 3 and 4 are normalized before being used in the loss functions.

Eventually, the overall objective function of the multi-task learning framework in this work would be the weighted sum of the terms in Equations 1, 2, 3 and 4:

$$O = L + \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3$$

where λ_1, λ_2 and λ_3 are the trade-off parameters. This concludes the presentation of the model in this work. We note that the proposed technique with transfer learning via connective and relation embeddings is general and can be applied on top of any multi-task learning models for IDRR to further improve their performance.

4 Experiments

4.1 Dataset, Resources and Parameters

We evaluate the models in this work on PDTB 2.0, one of the largest datasets that is commonly used for IDRR research. PDTB involves three levels of senses (relations): level 1 for 4 classes of relations, level 2 for 16 types of relations and level 3 for subtypes. We consider different settings for PDTB that have been studied in the previous research to achieve a comparable and comprehensive comparison, including the one-versus-others binary classifications for the first level (leading to four different datasets for the relations *Comparison*, *Contingency*, *Expansion* and *Temporal*), the muti-class classification setting for the first level (i.e., 4-way classification) and the multi-class classification for the second level (i.e., 11-way classification for the most popular types) (Pitler et al., 2009; Ji and Eisenstein, 2015b; Qin et al., 2017). Each setting has its own training, development and test datasets. For the 11-way classification setting, we further consider two popular ways to split the PDTB dataset, i.e., *PDTB-Lin* in (Lin et al., 2009) that use sections 2-21, 22 and 23 for the training, development and test datasets respectively, and *PDTB-Ji* (Ji and Eisenstein, 2015b; Bai and Zhao, 2018) where sections 2-20, 0-1, and 21-22 constitute the training, development and test datasets. In order to obtain the mapping between connectives and relations in the datasets, we rely on the association of the implicit connectives and relations in the examples of the training datasets.

We employ the same parameters and resources for the encoder model M as those in (Bai and Zhao, 2018) to achieve a fair comparison. We tune the dimension d of the relation and connective embeddings, and the trade-off parameters ($\lambda_1, \lambda_2, \lambda_3$) on the development sets for the aforementioned settings of the PDTB datasets, leading to $d = 80$ for different settings, and ($\lambda_1, \lambda_2, \lambda_3$) = (0.1, 0.01, 0.1), (0.01, 0.01, 0.01) and (1, 1, 1) for the binary, 4-way and 11-way classification experiments respectively. Following (Bai and Zhao, 2018), we use the Adagrad optimizer with learning rate = 0.001 to optimize the models in this work.

4.2 Comparing to the State of the Art

This section compares our proposed model with the current state-of-the-art models for IDRR. In particular, Table 1 reports the performance of the models for the one-versus-other binary classification settings while Table 2 shows the performance of the models for the multi-class classification settings (i.e., 4-way and 11-way with PDTB-Lin and PDTB-Ji) on the corresponding test sets.

System	Comp	Cont	Exp	Temp
(Pitler et al., 2009)	21.96	47.13	-	16.76
(Zhang et al., 2015)	33.22	52.04	69.59	30.54
(Chen et al., 2016)	40.17	54.76	-	31.32
(Qin et al., 2016b)	41.55	57.32	71.50	35.43
(Liu et al., 2016)	37.91	55.88	69.97	37.17
(Liu and Li, 2016b)	36.70	54.48	70.43	38.84
(Zhang et al., 2016)	35.88	50.56	71.48	29.54
(Qin et al., 2017)	40.87	54.56	72.38	36.20
(Lan et al., 2017)	40.73	**58.96**	72.47	38.50
(Dai and Huang, 2018)	46.79	57.09	70.41	**45.61**
(Lei et al., 2018)	43.24	57.82	72.88	29.10
(Guo et al., 2018)	40.35	56.81	72.11	38.65
(Bai and Zhao, 2018)	47.85	54.47	70.60	36.87
This work	**48.44**	56.84	**73.66**	38.60

Table 1: System performance (F1) for the binary classification settings.

System	4-way	PDTB-Lin	PDTB-Ji
(Lin et al., 2009)	-	40.20	-
(Ji and Eisenstein, 2015b)	-	-	44.59
(Qin et al., 2016)	-	43.81	45.04
(Liu and Li, 2016b)	46.29	-	-
(Qin et al., 2017)	-	44.65	46.23
(Lan et al., 2017)	47.80	-	-
(Dai and Huang, 2018)	51.84	-	-
(Lei et al., 2018)	47.15	-	-
(Guo et al., 2018)	47.59	-	-
(Bai and Zhao, 2018)	51.06	45.73	48.22
This work	**53.00**	**46.48**	**49.95**

Table 2: System performance for the multi-class classification settings (i.e., F1 for 4-way and Accuracy for PDTB-Lin and PDTB-Ji as in the prior work). Our model is significantly better than the others ($p < 0.05$).

The first observation from these tables is that the proposed model is significantly better than the model in (Bai and Zhao, 2018) over all the dataset settings (with $p < 0.05$) with large performance gap. As the proposed model is developed on top of the model in (Bai and Zhao, 2018), this is a direct comparison and demonstrates the benefit of the embeddings for relations and connectives as well as the transfer learning mechanisms for IDRR in this work. Second, the proposed model achieves the state-of-the-art performance on the multi-class classification settings (i.e., Table 2) and two settings for binary classification (i.e., *Comparison* and *Expansion*). The performance gaps between the proposed method and the other methods on the multi-class classification datasets (i.e., Table 2) are large and clearly testify to the advantage of the proposed model for IDRR.

4.3 Ablation Study

The multi-task learning framework in this work involves three penalization terms (i.e., L_1, L_2 and L_3 in Equations 2, 3 and 4). In order to illustrate the contribution of these terms, Table 3 presents the test set performance of the proposed model when different combinations of the terms are employed for the multi-class classification settings.

The row with "None" in the table corresponds to the proposed model where none of the penalization terms (L_1, L_2 and L_3) is used, reducing to the model in (Bai and Zhao, 2018) that is augmented with the connective and relation embeddings. As we can see from the table, the embeddings of connectives and relations can only slightly improve the performance of the model in (Bai and Zhao, 2018), necessitating the penalization terms L_1, L_2 and L_3 to facilitate the knowledge transfer and further improve the performance. From the table, it is also clear that each penalization term is important for the proposed model as eliminating any of them would worsen the performance. Combining the three penalization terms results in the best performance for IDRR in this work.

System	4-way	PDTB-Lin	PDTB-Ji
L1 + L2 + L3	**53.00**	**46.48**	**49.95**
L1 + L2	52.18	46.08	49.28
L1 + L3	52.31	45.30	49.57
L2 + L3	52.57	44.91	49.86
L1	51.11	46.21	49.09
L2	50.38	45.56	47.83
L3	52.52	45.69	49.09
None	51.62	45.82	48.60

Table 3: System performance with different combinations of $L1$, L_2 and L_3 (i.e., F1 for 4-way and Accuracy for PDTB-Lin and PDTB-Ji as in prior work). "None": not using any term.

5 Conclusion

We present a novel multi-task learning model for IDRR with deep learning. Our proposed model features the embeddings of the implicit connectives and discourse relations, and the three penalization terms to encourage the knowledge sharing between the prediction tasks. We achieve the state-of-the-art performance on different settings for the popular dataset PDTB for IDRR. In the future work, we plan to extend the idea of multi-task learning/transfer learning with label embeddings to the problems in information extraction (e.g., event detection, relation extraction, entity mention detection) (Nguyen and Grishman, 2015a,b, 2016d; Nguyen et al., 2016a,b,c; Nguyen and Nguyen, 2018b, 2019). In these problems, the labels are often organized in the hierarchies (e.g., types, subtypes) and the label embeddings can exploit such hierarchies to transfer the knowledge between different label-specific prediction tasks.

Acknowledgement

This research is partially supported by the Office of Naval Research Global (ONRG), Air Force Office of Scientific Research (AFOSR), and Asian Office of Aerospace Research & Development (AOARD) under Award Numbers N62909-18-1-2072 and 17IOA031.

References

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. In *COLING*.

Chloé Braud and Pascal Denis. 2015. Comparing word representations for implicit discourse relation classification. In *EMNLP*.

Deng Cai and Hai Zhao. 2017. Pair-aware neural sentence modeling for implicit discourse relation classification. In *IEA/AIE*.

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*.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *NAACL-HLT*.

Shima Gerani, Yashar Mehdad, Giuseppe Carenini, Raymond T. Ng, and Bita Nejat. 2014. Abstractive summarization of product reviews using discourse structure. In *EMNLP*.

Fengyu Guo, Ruifang He, Di Jin, Jianwu Dang, Longbiao Wang, and Xiangang Li. 2018. Implicit discourse relation recognition using neural tensor network with interactive attention and sparse learning. In *COLING*.

Peter Jansen, Mihai Surdeanu, and Peter Clark. 2014. Discourse complements lexical semantics for nonfactoid answer reranking. In *ACL*.

Yangfeng Ji and Jacob Eisenstein. 2015b. One vector is not enough: Entity-augmented distributed semantics for discourse relations. In *TACL*.

Yangfeng Ji, Gongbo Zhang, and Jacob Eisenstein. 2015a. Closing the gap: Domain adaptation from explicit to implicit discourse relations. In *EMNLP*.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2018. A knowledge-augmented neural network model for implicit discourse relation classification. In *COLING*.

Alistair Knott. 2014. A data-driven methodology for motivating a set of coherence relations. In *Ph.D. Thesis - The University of Edinburgh*.

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 *EMNLP*.

Man Lan, Yu Xu, and Zhengyu Niu. 2013. Leveraging synthetic discourse data via multi-task learning for implicit discourse relation recognition. In *ACL*.

Wenqiang Lei, Yuanxin Xiang, Yuwei Wang, Qian Zhong, Meichun Liu, and Min-Yen Kan. 2018. Linguistic properties matter for implicit discourse relation recognition: Combining semantic interaction, topic continuity and attribution. In *AAAI*.

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 *EMNLP*.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *ICRL*.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the penn discourse treebank. In *EMNLP*.

Yang Liu and Sujian Li. 2016b. Recognizing implicit discourse relations via repeated reading: Neural networks with multi-level attention. In *EMNLP*.

Yang Liu, Sujian Li, Xiaodong Zhang, and Zhifang Sui. 2016. Implicit discourse relation classification via multi-task neural networks. In *AAAI*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Minh Nguyen and Thien Huu Nguyen. 2018b. Who is killed by police: Introducing supervised attention for hierarchical lstms. In *COLING*.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016a. Joint event extraction via recurrent neural networks. In *NAACL*.

Thien Huu Nguyen, Lisheng Fu, Kyunghyun Cho, and Ralph Grishman. 2016c. A two-stage approach for extending event detection to new types via neural networks. In *Proceedings of the 1st ACL Workshop on Representation Learning for NLP (RepL4NLP)*.

Thien Huu Nguyen and Ralph Grishman. 2015a. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st NAACL Workshop on Vector Space Modeling for NLP (VSM)*.

Thien Huu Nguyen and Ralph Grishman. 2015b. Event detection and domain adaptation with convolutional neural networks. In *ACL-IJCNLP*.

Thien Huu Nguyen and Ralph Grishman. 2016d. Combining neural networks and log-linear models to improve relation extraction. In *Proceedings of IJCAI Workshop on Deep Learning for Artificial Intelligence (DLAI)*.

Thien Huu Nguyen, Avirup Sil, Georgiana Dinu, and Radu Florian. 2016b. Toward mention detection robustness with recurrent neural networks. In *Proceedings of IJCAI Workshop on Deep Learning for Artificial Intelligence (DLAI)*.

Trung Minh Nguyen and Thien Huu Nguyen. 2019. One for all: Neural joint modeling of entities and events. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT*.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *ACL-AFNLP*.

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

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. Implicit discourse relation recognition with context-aware character-enhanced embeddings. In *COLING*.

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

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *ACL*.

WenTing Wang, Jian Su, and Chew Lim Tan. 2010. Kernel based discourse relation recognition with temporal ordering information. In *EMNLP*.

Bonnie Webber, Alistair Knott, Matthew Stone, and Aravind Joshi. 1999. Discourse relations: A structural and presuppositional account using lexicalised tag. In *ACL*.

Changxing Wu, xiaodong shi, Yidong Chen, Yanzhou Huang, and jinsong su. 2016. Bilingually-constrained synthetic data for implicit discourse relation recognition. In *EMNLP*.

Changxing Wu, Xiaodong Shi, Yidong Chen, Jinsong Su, and Boli Wang. 2017. Improving implicit discourse relation recognition with discourse-specific word embeddings. In *ACL*.

Yasuhisa Yoshida, Jun Suzuki, Tsutomu Hirao, and Masaaki Nagata. 2014. Dependency-based discourse parser for single-document summarization. In *EMNLP*.

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*.

Biao Zhang, Deyi Xiong, jinsong su, Qun Liu, Rongrong Ji, Hong Duan, and Min Zhang. 2016. Variational neural discourse relation recognizer. In *EMNLP*.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *COLING*.