Next Sentence Prediction helps Implicit Discourse Relation Classification within and across Domains

Wei Shi[†] and **Vera Demberg**[†,‡]
[†]Dept. of Language Science and Technology
[‡]Dept. of Mathematics and Computer Science, Saarland University
Saarland Informatics Campus, 66123 Saarbrücken, Germany
{w.shi, vera}@coli.uni-saarland.de

Abstract

Implicit discourse relation classification is one of the most difficult tasks in discourse parsing. Previous studies have generally focused on extracting better representations of the relational arguments. In order to solve the task, it is however additionally necessary to capture what events are expected to cause or follow each other. Current discourse relation classifiers fall short in this respect. We here show that this shortcoming can be effectively addressed by using the bidirectional encoder representation from transformers (BERT) proposed by Devlin et al. (2019), which were trained on a next-sentence prediction task, and thus encode a representation of likely next sentences. The BERT-based model outperforms the current state of the art in 11-way classification by 8% points on the standard PDTB dataset. Our experiments also demonstrate that the model can be successfully ported to other domains: on the BioDRB dataset, the model outperforms the state of the art system around 15% points.

1 Introduction

Discourse relation classification has been shown to be beneficial to multiple down-stream NLP tasks such as machine translation (Li et al., 2014), question answering (Jansen et al., 2014) and summarization (Yoshida et al., 2014). Following the release of the Penn Discourse Tree Bank (Prasad et al., 2008, PDTB), discourse relation classification has received a lot of attention from the NLP community, including two CoNLL shared tasks (Xue et al., 2015, 2016).

Discourse relations in texts are sometimes marked with an explicit connective (e.g., *but*, *because*, *however*), but these explicit signals are often absent. With explicit connectives acting as informative cues, it is relatively easy to classify the discourse relation with high accuracy (93.09% on four-way classification in (Pitler et al., 2008)). When there is no connective, classification has to rely on semantic information from the relational arguments. This task is very challenging, with state-of-the-art systems achieving accuracy of only 45% to 48% on 11-way classification. Consider example 1:

(1) [*The joint venture with Mr. Lang wasn't a good one.*]$_{Arg1}$ [**The venture, formed in 1986, was supposed to be Time's low-cost, safe entry into women's magazines.**]$_{Arg2}$
implicit `Comp.Concess.expectation` relation from PDTB: wsj_1903

In order to correctly classify the relation, it is necessary to understand that Arg1 raises the expectation that the next discourse segment may provide an explanation for why the venture wasn't good (e.g., that it was risky), and Arg2 contrasts with this discourse expectation. More generally, this means that a successful discourse relation classification model would have to be able to learn typical temporal event sequences, reasons, consequences etc. for all kinds of events. Statistical models attempted to address this intuition by giving models word pairs from the two arguments as features (Lin et al., 2009; Park and Cardie, 2012; Biran and McKeown, 2013; Rutherford and Xue, 2014), so that models could for instance learn to recognize antonym relations between words in the two arguments.

Recent models exploit such similarity relations between the two arguments, as well as simpler surface features that occur in one relational argument and correlate with specific coherence relations (e.g., the presence of negation, temporal expressions etc. may give hints as to what coherence relation may be present, see Park and Cardie (2012); Asr and Demberg (2015)). However, relations between arguments are often a lot more diverse than simple contrasts that can be captured

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5790–5796, Hong Kong, China, November 3–7, 2019.

through antonyms, and may rely on world knowledge (Kishimoto et al., 2018). It is hence clear that one cannot learn all these diverse relations from the very small amounts of available training data. Instead, we would have to learn a more general representation of discourse expectations.

Many recent discourse relation classification approaches have focused on cross-lingual data augmentation (Shi et al., 2017, 2019), training models to better represent the relational arguments by using various neural network models, including feed-forward network (Rutherford et al., 2017), convolutional neural networks (Zhang et al., 2015), recurrent neural network (Ji et al., 2016; Bai and Zhao, 2018), character-based (Qin et al., 2016) or formulating relation classification as an adversarial task (Qin et al., 2017). These models typically use pre-trained semantic embeddings generated from language modeling tasks, like Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014) and ELMo (Peters et al., 2018).

However, previously proposed neural models still crucially lack a representation of the *typical relations between sentences*: to solve the task properly, a model should ideally be able to form discourse expectations, i.e., to represent the typical causes, consequences, next events or contrasts to a given event described in one relational argument, and then assess the content of the second relational argument with respect to these expectations (see Example 1). Previous models would have to learn these relations only from the annotated training data, which is much too sparse for learning all possible relations between all events, states or claims.

The recently proposed BERT model (Devlin et al., 2019) takes a promising step towards addressing this problem: the BERT representations are trained using a language modelling and, crucially, a "next sentence prediction" task, where the model is presented with the actual next sentence vs. a different sentence and needs to select the original next sentence. We believe it is a good fit for discourse relation recognition, since the task allows the model to represent what a typical next sentence would look like.

In this paper, we show that a BERT-based model outperforms the current state of the art by 8% points in 11-way implicit discourse relation classification on PDTB. We also show that after pre-trained with small size cross-domain data, the model can be easily transferred to a new domain: it achieves around 16% accuracy gain on BioDRB compared to state of the art model. We also show that the Next Sentence Prediction task played an important role in these improvements.

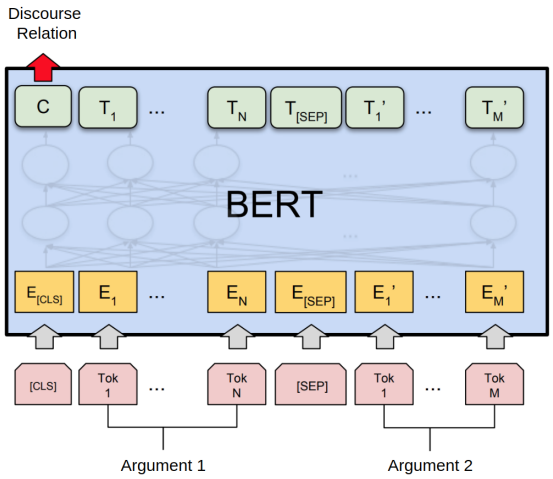

Figure 1: The architecture from BERT (Devlin et al., 2019) for fine-tuning of implicit discourse relation classification.

2 Next Sentence Prediction

Devlin et al. (2019) proposed the Bidirectional Encoder Representation from Transformers (BERT), which is designed to pre-train a deep bidirectional representation by jointly conditioning on both left and right contexts. BERT is trained using two novel unsupervised prediction tasks: Masked Language Modeling and Next Sentence Prediction (NSP). The NSP task has been formulated as a binary classification task: the model is trained to distinguish the original following sentence from a randomly chosen sentence from the corpus, and it showed great helps in multiple NLP tasks especially inference ones. The resulting BERT representations thus encode a representation of upcoming discourse content, and hence contain discourse expectation representations which, as we argued above, are required for classifying coherence relations.

3 Experiments and Results

As shown in Figure 1, E denotes the input tokens' embedding and T are the target words. In our case they are identical. We fit the implicit discourse relation task into sentence-pair classification proposed in BERT. Argument 1 and Argu-

Methods	PDTB-Lin	PDTB-Ji	Cross Validation
Cai and Zhao (2017)	-	45.81	-
Kishimoto et al. (2018)	38.77	-	39.80
Bai and Zhao (2018)	45.73	48.22	-
Shi and Demberg (2019)	45.82	47.83	41.29
Bi-LSTM + w2v_300	37.95(0.59)	40.57(0.67)	37.82(0.74)
BERT	**53.13(0.37)**	**53.30(0.39)**	**49.30(1.33)**
BERT + WSJ w/o NSP	**53.39(0.49)**	**51.28(0.49)**	**49.32(1.24)**
BERT + WSJ	**54.82(0.61)**$*$	**53.23(0.32)**$*$	**49.35(0.83)**

Table 1: Accuracy (%) with standard deviation in brackets of implicit discourse relation classification on different settings of PDTB level 2 relations. NSP refers to the subtask "next sentence prediction" in the pre-training of BERT. Numbers in bold signal significant improvements over the previous state of the art ($p<0.01$). Numbers with $*$ denote significant improvements over BERT + WSJ w/o NSP with $p<0.01$.

ment 2 are separated with token "[SEP]"; "[CLS]" is the special classification embedding while "C" is the same as "[CLS]" in pre-training but the ground-truth label in the fine-tuning. In the experiments, we used the uncased base model[1] provided by Devlin et al. (2019), which is trained on *BooksCorpus* and *English Wikipedia* with 3300M tokens in total.

3.1 Evaluation on PDTB

We used the Penn Discourse Tree Bank (Prasad et al., 2008), the largest available manually annotated discourse corpus. It provides a three level hierarchy of relation tags. Following the experimental settings and evaluation metrics in Bai and Zhao (2018), we use two most-used splitting methods of PDTB data, denoted as PDTB-Lin (Lin et al., 2009), which uses sections 2-21, 22, 23 as training, validation and test sets, and PDTB-Ji (Ji and Eisenstein, 2015), which uses 2-20, 0-1, 21-22 as training, validation and test sets and report the overall accuracy score. In addition, we also performed 10-fold cross validation among sections 0-22, as promoted in Shi and Demberg (2017). We also follow the standard in the literature to formulate the task as an 11-way classification task.

Results are presented in Table 1. We evaluated three versions of the BERT-based model. All of our BERT models use the pre-trained representations and are fine-tuned on the PDTB training data. The version marked as "BERT" does not do any additional pre-training. BERT+WSJ in addition performs further pre-training on the parts of the *Wall Street Journal* corpus that do not have discourse relation annotation. The model version "BERT+WJS w/o NSP" also performs pre-training on the WSJ corpus, but only uses the Masked Language Modelling task, not the Next Sentence Prediction task in the pre-training. We added this variant to measure the benefit of in-domain NSP on discourse relation classification (note though that the downloaded pre-trained BERT model contains the NSP task in the original pre-training).

We compared the results with four state-of-the-art systems: Cai and Zhao (2017) proposed a model that takes a step towards calculating discourse expectations by using attention over an encoding of the first argument, to generate the representation of the second argument, and then learning a classifier based on the concatenation of the encodings of the two discourse relation arguments. Kishimoto et al. (2018) fed external world knowledge (ConceptNet relations and coreferences) explicitly into MAGE-GRU (Dhingra et al., 2017) and achieved improvements compared to only using the relational arguments. However, we here show that it works even better when we learn this knowledge implicit through next sentence prediction task. Shi and Demberg (2019) used a seq2seq model that learns better argument representations due to being trained to explicitate the implicit connective. In addition, their classifier also uses a memory network that is intended to help remember similar argument pairs encountered during training. The current best performance was achieved by Bai and Zhao (2018), who combined representations from different grained em-

[1] https://github.com/google-research/bert#pre-trained-models

beddings including contextualized word vectors from ELMo (Peters et al., 2018), which has been proved very helpful. In addition, we compared our results with a simple bidirectional LSTM network and pre-trained word embeddings from Word2Vec.

We can see that on all settings, the model using BERT representations outperformed all existing systems with a substantial margin. It obtained improvements of 7.3% points on PDTB-Lin, 5.5% points on PDTB-Ji, compared with the ELMo-based method proposed in (Bai and Zhao, 2018). What's more, the BERT model outperformed (Shi and Demberg, 2019) on cross validation by around 8%, with significance of $p<0.01$. Significance test was performed by estimating variance of the model from the performance on different folds in cross-validation (paired t-test). For the Lin and Ji evaluations, we estimated variance due to random initialization by running them 5 times and calculating the likelihood that the state-of-the-art model result would come from that distribution.

3.2 Evaluation On BioDRB

The Biomedical Discourse Relation Bank (Prasad et al., 2011) also follows PDTB-style annotation. It is a corpus annotated over 24 open access full-text articles from the GENIA corpus (Kim et al., 2003) in the biomedical domain. Compared with PDTB, some new discourse relations and changes have been introduced in the annotation of BioDRB. In order to make the results comparable, we preprocessed the BioDRB annotations to map the relations to the PDTB ones, following the instructions in Prasad et al. (2011).

The biomedical domain is very different from the WSJ or the data on which the BERT model was trained. The BioDRB contains a lot of professional words / phrases that are extremely hard to model. In order to test the ability of the BERT model on cross-domain data, we performed fine-tuning on PDTB while testing on BioDRB. We also tested the state of the art model of implicit discourse relation classification proposed by Bai and Zhao (2018) on BioDRB. From Table 2, we can see that the BERT base model achieved almost 12% points improvement over the Bi-LSTM baseline and 15% points over Bai and Zhao (2018). When fine-tuned on in-domain data in the cross-validation setting, the improvement increases to around 17% points.

Method	Cross-Domain	In-Domain
Bi-LSTM + w2v_300	32.97	46.49
Bai and Zhao (2018)	29.52	55.90
BioBERT (Lee et al., 2019)	**44.33**	**67.58**
BERT	**44.79**	**63.02**
BERT + GENIA w/o NSP	**43.99**	**65.02**
BERT + GENIA	**45.19**$*$	**66.04**$*$

Table 2: Accuracy (%) on BioDRB level 2 relations with different settings. Cross-Domain means trained on PDTB and tested on BioDRB. For the In-Domain setting, we used 5-fold cross-validation and report average accuracy. Numbers in bold are significantly better than the state of the art system with $p<0.01$ and numbers with $*$ denote denote significant improvements over BERT + GENIA w/o NSP with $p<0.01$.

It is also interesting to know whether the performance of the BERT model can be improved if we add additional pre-training on in-domain data. BioBert (Lee et al., 2019) continues pre-training BERT with bio-medical texts including PubMed and PMC corpora (around 18B tokens), which achieved the best results on in-domain setting. Similarly, BERT+GENIA refers to a model in which the downloaded BERT representations are further pre-trained on the parts of the GENIA corpus which consists of 18k sentences and is not annotated with coherence relations. Evaluation shows that this in-domain pre-training yields another 3% point improvement; our tests also show that the NSP task again plays a substantial role in the improvement. We believe that gains for further pre-training on GENIA for the biomedical domain are higher than for pre-training on WSJ for PDTB because the domain difference between the *BooksCorpus* and the biomedical domain is larger.

Currently there are not so many published results that we can compare with on BioDRB for implicit discourse relation classification. We compared BERT model with naïve Bayes and MaxEnt methods proposed in Xu et al. (2012) on one-versus-all binary classification. We followed the settings in Xu et al. (2012) and used two articles ("GENIA_1421503", "GENIA_1513057") for testing and one article ("GENIA_111020") for validation. During training, we employed down-sampling or up-sampling to keep the numbers of positive and negative samples in each relation consistent. The BERT base model achieved 43.03% average F_1 score and 77.34% average accuracy in one-versus-all level-1 classification. Compared with the current state-of-the-art perfor-

Method		Comp.	Cont.	Exp.	Temp.	Average
Naïve Bayes	(Xu et al., 2012)	7.69(61.45)	3.88(60.24)	68.54(55.02)	17.19(57.43)	24.33(58.54)
MaxEnt	(Xu et al., 2012)	7.08(57.83)	3.36(53.82)	72.32(60.64)	23.81(61.45)	26.64(58.44)
BERT		**8.25(63.82)**	**10.26(71.54)**	**90.20(86.18)**	**63.41(87.80)**	**43.03(77.34)**

Table 3: F_1-score (Accuracy) of binary classification on level 1 implicit relation in BioDRB.

Relations	WSJ w/o NSP			WSJ w/ NSP			C.	GENIA w/o NSP			GENIA w/ NSP			C.
	P	R	F_1	P	R	F_1		P	R	F_1	P	R	F_1	
Asynchronous	0.38	0.46	0.41	0.29	0.38	0.33	13	-	-	-	-	-	-	-
Synchrony	-	-	-	-	-	-	5	0.87	0.84	0.85	0.90	0.88	0.89	80
Cause	0.57	0.65	0.61	0.57	0.64	0.60	200	0.22	0.10	0.14	0.23	0.15	0.18	20
Prag. Cause	-	-	-	-	-	-	5	-	-	-	-	-	-	1
Contrast	0.55	0.48	0.51	0.54	0.57	0.55	127	-	-	-	-	-	-	22
Concession	-	-	-	-	-	-	5	-	-	-	0.50	0.06	0.11	16
Conjunction	0.42	0.60	0.49	0.46	0.61	0.53	118	0.60	0.78	0.68	0.62	0.82	0.71	130
Instantiation	0.62	0.67	0.64	0.62	0.65	0.64	72	-	-	-	-	-	-	9
Restatement	0.52	0.45	0.48	0.55	0.45	0.50	190	0.56	0.76	0.65	0.59	0.69	0.64	72
Alternative	0.83	0.33	0.48	0.67	0.40	0.50	15	-	-	-	-	-	-	1
List	0.71	0.17	0.27	0.60	0.20	0.30	30	-	-	-	-	-	-	-
Macro Avg.	0.53	0.53	0.52	0.55	0.55	0.55	780	0.55	0.64	0.59	0.59	0.66	0.61	351

Table 4: Precision, Recall and F_1 score for each level-2 relation on PDTB-Lin setting and BioDRB with "BERT + WSJ/GENIA" systems w/ and w/o NSP. "-" indicates 0.00 and "C." means the number of each relation in the test set.

mances (26.64% F_1 and 58.54% accuracy) in Xu et al. (2012), it achieves around 16% and 19% points improvement when trained in-domain, as illustrated in Table 3.

3.3 Discussion

The usage of the BERT model in this paper was motivated primarily by the use of the next-sentence prediction task during training. The results in Table 1 and Table 2 confirm that removing the "Next Sentence Prediction" hurts the performance on both PDTB and BioDRB.

In order to have better insights about which relation has benefited from the NSP task, we also reported the detailed performance for each relation with and without it in BERT. As illustrated in Table 4, we can see that performances on relations like *Temporal.Synchrony, Comparison.Contrast, Expansion.Conjunction and Expansion.Alternative* have been improved by a large margin. This shows that representing the likely upcoming sentence helps the model form discourse expectations, which the classifier can then use to predict the coherence relation between the actually observed arguments.

However, compared with BERT+GENIA, the results of BioBert (Lee et al., 2019) in Table 2 show that having large in-domain data for pre-training also has limited ability in learning domain specific representations. We therefore believe that the model could be further improved by including external domain-specific knowledge from an ontology (as in Kishimoto et al. (2018)) or a causal graph for biomedical concepts and events.

4 Conclusion and Future work

In this paper, we show that BERT has very good ability in encoding the semantic relationship between sentences with its "next sentence prediction" task in pre-training. It outperformed the current state-of-the-art systems significantly with a substantial margin on both in-domain and cross domain data. Our results also indicate that the next-sentence prediction task during training indeed plays a role in this improvement. Future work should explore the joint representation of discourse expectations through implicit representations that are learned during training and the inclusion of external knowledge. In addition, Yang et al. (2019) showed that NSP only helps tasks with longer texts. It would be interesting to see whether it has the same effect on implicit discourse relation classification task, we'd like to leave that in the future work.

5 Acknowledgements

This research was funded by the German Research Foundation (DFG) as part of SFB 1102 "Information Density and Linguistic Encoding". We'd like to thank all the reviewers for their insightful and valuable comments.

References

Fatemeh Torabi Asr and Vera Demberg. 2015. Uniform surprisal at the level of discourse relations: Negation markers and discourse connective omission. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 118–128.

Hongxiao Bai and Hai Zhao. 2018. Deep enhanced representation for implicit discourse relation recognition. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 571–583.

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*, pages 69–73, Sofia, Bulgaria. Association for Computational Linguistics.

Deng Cai and Hai Zhao. 2017. Pair-aware neural sentence modeling for implicit discourse relation classification. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 458–466. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Bhuwan Dhingra, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. 2017. Linguistic knowledge as memory for recurrent neural networks. *arXiv preprint arXiv:1703.02620*.

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 (Volume 1: Long Papers)*, volume 1, pages 977–986.

Yangfeng Ji and Jacob Eisenstein. 2015. One vector is not enough: Entity-augmented distributed semantics for discourse relations. *Transactions of the Association of Computational Linguistics*, 3(1):329–344.

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.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. 2003. Genia corpusa semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Yudai Kishimoto, Yugo Murawaki, and Sadao Kurohashi. 2018. A knowledge-augmented neural network model for implicit discourse relation classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 584–595.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

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

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*, pages 343–351. Association for Computational Linguistics.

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*, pages 3111–3119.

Joonsuk Park and Claire Cardie. 2012. Improving implicit discourse relation recognition through feature set optimization. In *13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 108.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind Joshi. 2008. Easily identifiable discourse relations. In *Coling 2008: Companion volume: Posters*, pages 87–90.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Rashmi Prasad, Susan McRoy, Nadya Frid, Aravind Joshi, and Hong Yu. 2011. The biomedical discourse relation bank. *BMC bioinformatics*, 12(1):188.

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

Lianhui Qin, Zhisong Zhang, Hai Zhao, Zhiting Hu, and Eric Xing. 2017. Adversarial connective-exploiting networks for implicit discourse relation classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1006–1017.

Attapol Rutherford, Vera Demberg, and Nianwen Xue. 2017. A systematic study of neural discourse models for implicit discourse relation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 281–291.

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*, pages 645–654. Association for Computational Linguistics.

Wei Shi and Vera Demberg. 2017. Do we need cross validation for discourse relation classification? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 150–156.

Wei Shi and Vera Demberg. 2019. Learning to explicitate connectives with seq2seq network for implicit discourse relation classification. In *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pages 188–199.

Wei Shi, Frances Yung, and Vera Demberg. 2019. Acquiring annotated data with cross-lingual explicitation for implicit discourse relation classification. In *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, pages 12–21.

Wei Shi, Frances Yung, Raphael Rubino, and Vera Demberg. 2017. Using explicit discourse connectives in translation for implicit discourse relation classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 484–495.

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

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. Association for Computational Linguistics.

Nianwen Xue, Hwee Tou Ng, 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. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

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.

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*, pages 2230–2235.