On the Need of Cross Validation for Discourse Relation Classification

Wei Shi
Dept. of Language Science and Technology
Saarland University
66123 Saarbrücken, Germany
w.shi@coli.uni-saarland.de

Vera Demberg
Saarland Informatics Campus
Saarland University
66123 Saarbrücken, Germany
vera@coli.uni-saarland.de

Abstract

The task of implicit discourse relation classification has received increased attention in recent years, including two CoNNL shared tasks on the topic. Existing machine learning models for the task train on sections 2-21 of the PDTB and test on section 23, which includes a total of 761 implicit discourse relations. In this paper, we'd like to make a methodological point, arguing that the standard test set is too small to draw conclusions about whether the inclusion of certain features constitute a genuine improvement, or whether one got lucky with some properties of the test set, and argue for the adoption of cross validation for the discourse relation classification task by the community.

1 Introduction

Discourse-level relation analysis is relevant to a variety of NLP tasks such as summarization (Yoshida et al., 2014), question answering (Jansen et al., 2014) and machine translation (Meyer et al., 2015). Recent years have seen more and more works on this topic, including two CoNNL shared tasks (Xue et al., 2015; Xue et al., 2016). The community most often uses the Penn Discourse Treebank (PDTB) (Prasad et al., 2008) as a resource, and has adopted the usual split into training and test data as used for other tasks such as parsing. Because discourse relation annotation is at a higher level than syntactic annotation, this however means that the test set is rather small, and with the amount of alternative features and, more recently, neural network architectures being applied to the problem, we run a serious risk as a community of believing in features that are successful in getting some improvement on the specific test set but don't generalize at all.

In discourse relation parsing, we usually distinguish between *implicit* and *explicit* discourse relations. Explicit relations are marked with a discourse connective such as "because", "but", "if", while implicit discourse relations are not marked with any discourse connective. The connective serves as a strong cue for the discourse relation, as the example below demonstrates:

*" Typically, money-fund yields beat comparable short-term investments **because** portfolio managers can vary maturities and go after the highest rates" (Explicit, Contingency.Cause)*

*" They desperately needed somebody who showed they cared for them, who loved them. (**But**) The last thing they needed was another drag-down blow." (Implicit, Comparison.Contrast)*

Previous studies show that the presence of connectives can greatly help with classification of the relation and can be disambiguated with 0.93 accuracy (4-ways) solely on the discourse relation connectives (Pitler et al., 2008). In implicit relations, no such strong cue is available and the discourse relation instead needs to be inferred based on the two textual arguments.

In recent studies, various classes of features are explored to capture lexical and semantic regularities for identifying the sense of implicit relations, including linguistically informed features like polarity tags, Levin verb classes, length of verb phrases, language model based features, contextual features, constituent parse features and dependency parse features (Lin et al., 2009; Pitler et al., 2009; Zhou et al., 2010; Zhang et al., 2015; Chen et al., 2016). For some of second-level relations (a level of granularity that should be much more meaningful to downstream tasks than the four-way distinction), there are only a dozen in-

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 150–156,
Valencia, Spain, April 3-7, 2017.

stances, so that it's important to make maximal use of both the data set for training and testing. The test set that is currently most often used for 11 way classification is section 23 (Lin et al., 2009; Ji and Eisenstein, 2015; Rutherford et al., 2017), which contains only about 761 implicit relations. This small size implies that a gain of 1 percentage point in accuracy corresponds to just classifying an additional 7-8 instances correctly.

This paper therefore aims to demonstrate the degree to which conclusions about the effectiveness of including certain features would depend on whether one evaluates on the standard test section only, or performs cross validation on the whole dataset for second-level discourse relation classification. The model that we use is a neural network that takes the words occurring in the relation arguments as input, as well as traditional features mentioned above, to make comparisons with most-used section splits. To our knowledge, this is the first paper that systematically evaluates the effect of the train/test split for the implicit discourse relation classification task on PDTB. We report the classification performances on random and conventional split sections.

As a model, we use a neural network that also includes some of the surface features that have been shown to be successful in previous work. Our model is competitive with the state of the art. The experiments here are exemplary for what kind of conclusions we would draw from the cross validation vs. from the usual train-test split. We find that results are quite different in the different splits of dataset, which we think is a strong indication that cross validation is important to adopt as a standard practice for the discourse relation classification community. We view cross validation as an important method in case other unseen datasets are not available (note that at least for English, new datasets have recently been made available as part of the shared task (Xue et al., (2015; 2016); as well as Rehbein et al., (2016)).

2 Background on Discourse Relation Parsing

Soricut and Marcu (2003) firstly addressed the task of parsing discourse structure within the same sentence. Many of the useful features proposed by them, syntax in particular, revealed that both arguments of the connectives are found in the same sentence. The release of PDTB, the largest available annotated corpora of discourse relations, opened the door to machine learning based discourse relation classification.

Feature-based methods exploit discriminative features for implicit relation classification. Pitler et al. (2009) demonstrated that features developed to capture word polarity, verb classes and orientation, as well as some lexical features are strong indicator of the type of discourse relation. Lin et al. (2009) further introduced contextual, constituent and dependency parse features. They achieved an accuracy of 40.2% for 11-way classification, a 14.1% absolute improvement over the baseline. With these features, Park and Cardie (2012) provided a systematic study of previously proposed features and identified feature combinations. Additional features proposed later include relation specific word similarity (Biran and McKeown, 2013), Brown clusters and Coreference Patterns (Rutherford and Xue, 2014).

Data selection and extension is another main aspect for discourse relation classification, given that the number of training instances is limited and only from a single domain. Wang et al. (2012) proposed a novel single centroid clustering algorithm to differentiate typical and atypical examples for each discourse relation. Mihil et al. (2014) and Hernault et al. (2010) proposed semi-supervised learning methods to recognise relations. Rutherford and Xue (2015) collected additional training data from unannotated data, selecting instances based on two criteria (the degree to which a connective can generally be omitted and the degree to which a connective typically changes the interpretation of the relation) improved the inference of implicit discourse relation. Hidey and McKeown (2016), Quirk and Poon (2016) extended training data with weakly labeled data which are cheaply obtained by distant-supervised learning.

Recently the distributed word representations (Bengio et al., 2003; Mikolov et al., 2013) have shown an advantage in dealing with data sparsity problem (Braud and Denis, 2015). Many deep learning methods have been proved to be helpful in discourse relation parsing and achieved some significant progresses. Zhang et al. (2015) proposed a shallow convolutional neural network for implicit discourse recognition to alleviate the overfitting problem and help preserve the recognition and generalization ability with the model. Ji et al. (2015) computed distributed meaning represen-

tations for each discourse argument with recursive neural network. Ji et al. (2016) introduced a latent variable to recurrent neural network and outperformed in two tasks. Chen et al. (2016) adopted a gated relevance network to capture the semantic interaction between word pairs. Zhang et al. (2016) proposed a neural discourse relation recognizer with a semantic memory and attention weights for implicit discourse relation recognition.

The model we use in this paper is most closely related to the neural network model proposed in Rutherford et al. (2017). The model also has access to the traditional features, which are concatenated to the neural representations of the arguments in the output layer. In order to simulate what conclusions we would be drawing from comparing the contributions of the handcrafted surface features, we calculate accuracy for each of the handcrafted features.

3 Corpora

The Penn Discourse Treebank (PDTB) We use the Penn Discourse Treebank (Prasad et al., 2008), the largest available manually annotated corpora of discourse on top of one million word tokens from the Wall Street Journal (WSJ). The PDTB provides annotations for explicit and implicit discourse relations. By definition, an explicit relation contains an explicit discourse connective while the implicit one does not. The PDTB provides a three level hierarchy of relation tags for its annotation. Previous work in this task has been done over two schemes of evaluation: first-level 4-ways classification (Pitler et al., 2009; Rutherford and Xue, 2014; Chen et al., 2016), second-level 11-way classification (Lin et al., 2009; Ji and Eisenstein, 2015). The distribution of second-level relations in PDTB is illustrated in Table 1.

We follow the preprocessing method in (Lin et al., 2009; Rutherford et al., 2017). If the instance is annotated with two relations, we adopt the first one shown up, and remove those relations with too few instances. We treat section 2-21 as training set, section 22 as development set and section 23 as test set for our results reported as "most-used split". In order to investigate whether the results for benefit of including a certain feature to the model are stable, we conduct 10-fold cross-validation on the whole corpus including sections 0-24. Note that we here included also the validation section for our experiments, to have maximal data for our demonstration of variability between folds. For best practice when testing new models, we instead recommend to keep the validation set completely separate and do cross-validation for the remaining data. Also note that you might want to choose repeated cross-validation (which simply repeats the cross-validation step several times with the data divided up into different folds) as an alternative to simple cross-validation performed here. For a more in-detail discussion of cross validation methods, see (Kim, 2009; Bengio and Grandvalet, 2005).

In Table 1, we can see that the different relations' proportions on the training and test set are quite different in the most-used split. For instance, temporal relations are under-represented which may lead to a misestimation of the usefulness of features that are relevant for classifying temporal relations. For our cross validation experiments, we evenly divided all the instances in section 0-24 into 10 balanced folds[1]. The proportions of each class in the training and testing set are identical. With the same distribution of each class, we here avoid having an unbalanced number of instances per class among training and testing set.

4 Model

The task is to predict the discourse relation given the two arguments of an implicit instance. As a label set, we use 11-way distinction as proposed in Lin et al., (2009); Ji and Eisenstein (2015). Word Embeddings are trained with the Skip-gram architecture in *Word2Vec* (Mikolov et al., 2013), which is able to capture semantic and syntactic patterns with an unsupervised method, on the training sections of WSJ data.

Our model is illustrated in Figure 1. Each word is represented as a vector, which is found through a look-up word embedding. Then we get the representations of argument 1 and argument 2 separately after transforming semantic word vectors into distributed continuous-value features by LSTM recurrent neural network. With concatenating feature vector and the instance's representation, we classify it with a softmax layer and output its label.

Implementation All the models are implemented

[1]While we here chose balanced distributions, other designs of splitting up the data into folds such that different folds have organically different distributions of classes can alternatively be argued for, on the basis of more accurately representing new in-domain data distributions.

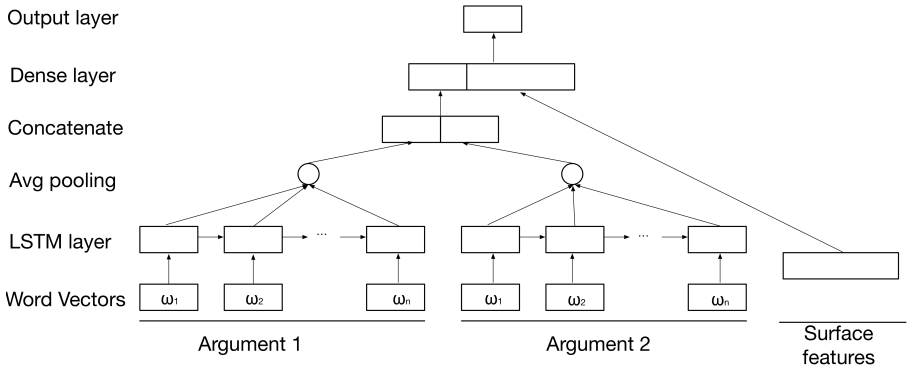

Figure 1: Long Short-Term Memory Model with surface features.

Relation	Most-used Split				Cross Validation *	
	Train		Test		Train	Test
Temporal.Asynchronous	542	(4.25%)	12	(1.58%)	583	65
Temporal.Synchrony	150	(1.18%)	5	(0.66%)	155	18
Contingency.Cause	3259	(25.53%)	193	(25.36%)	3581	398
Contingency.Pragmatic cause	55	(0.43%)	5	(0.66%)	61	7
Comparison.Contrast	1600	(12.54%)	126	(16.56%)	1843	205
Comparison.Concession	189	(1.48%)	5	(0.66%)	194	22
Expansion.Conjunction	2869	(22.48%)	116	(15.24%)	3075	342
Expansion.Instantiation	1130	(8.85%)	69	(9.07%)	1254	140
Expansion.Restatement	2481	(19.44%)	190	(24.97%)	2792	311
Expansion.Alternative	151	(1.18%)	15	(1.97%)	160	18
Expansion.List	337	(2.64%)	25	(3.29%)	347	39
Total	12763		761		14045	1565

* Numbers are averaged over different folds

Table 1: The distribution of training and test sets in Most-used Split and Cross Validation on level 2 relations in PDTB. Five types that have only have very few training instances are removed.

	Models	Most-used Split	Cross Validation
	Most common class	25.36	25.59
	Lin et al. (2009)	40.20	-[1]
	Ji & Eisenstein (2015) (surface features only)	40.66	-
	Rutherford et al. (2017)	39.56	-
Neural Network	No additional surface features	37.68	34.44 ($\pm$1.37)
Neural Network	Inquirer Tags	40.46	33.58 ($\pm$1.36) (2+,8-)
Neural Network	BrownCluster	38.77	33.83 ($\pm$1.59) (3+,7-)
Neural Network	Levin Class	40.92	34.17 ($\pm$1.48) (4+,6-)
Neural Network	Verbs	40.21	34.26 ($\pm$1.22) (5+,5-)
Neural Network	Modality	40.82	37.65 ($\pm$1.83) (6+,4-)
Neural Network	All Features above	38.56	35.90 ($\pm$1.32) (2+,8-)

[1] "-" means no result currently.

Table 2: Performance comparison of different features in Most-used Split and Cross Validation on second-level relations. Numbers for cross validation indicate the mean accuracy across folds, the standard deviation, and the number of folds that show better vs. worse performance when including the feature.

in Keras[2], which runs on top of Theano. The architecture of the model we use is illustrated in Figure 1. Regarding the initialization, regularization and learning algorithm, we follow all the settings in (Rutherford et al., 2017). We adopt cross-entropy as our cost function, adagrad as the optimization algorithm, initialized all the weights in the model with uniform random and set dropout layers after the embedding and output layer with a drop rate of 0.2 and 0.5 respectively.

5 Features

For the sake of our cross-validation argument, we choose five kinds of most popular features in discourse relation classification, namely *Inquirer Tags* (semantic classification tags), *Brown Clusters*, *Verb* features, *Levin classes* and *Modality*.

6 Results

We tested five frequently-used surface features with our model. Results are shown in Table 2. We can see that our implemented model is comparable with state of the art models. Our main point here is however not to argue that we outperform any particular model, but rather we'd like to discuss what conclusions we'd be drawing from adding surface features to our NN model if using the standard test set vs. doing cross validation.

For each cross validation with different features, the separation into train and test sets are identical. We can see that the performances on Most-used Split section is generally 3-7% better than the results for the rest of the corpus. While we would also conclude from our model when evaluated on the standard test set that each of these features contribute some useful information, we can also see that we would come to very different conclusions if actually running the cross-validation experiment.

Cross Validation is primarily a way of measuring the predictive performance of a model. With such a small test set, improvements on the classification could be the results of many factors. For instance, take a look at the effectiveness of including Inquirer Tags: these lead to an increase in performance by 2.8% in Most-used Split, but actually only helped on two out of 10-fold in the cross-validation set, overall leading to a small decrease in performance of the classifier. Similarly, the verb features seem to indicate a substantial improvement in relation classification accuracy on the standard test set, but there is no effect at all across the folds.

Other works, such as Berg-Kirkpatrick et al. (2012) strongly recommend significance testing to validate metric gains in NLP tasks, even though the relationship between metric gain and statistical significance is complex. We observed that recent papers in discourse relation parsing do not always perform significance testing, and if they do report significance, then oftentimes they do not report the test that was used. We would here like to argue in favour of significance testing with cross validation, as opposed to boot strapping methods that only use the standard test set. Due to the larger amount of data, calculating significance based on the cross validation will give us substantially better estimates about the robustness of our results, because it can quantify more exactly the amount of variation with respect to transferring to a new (in-domain) dataset.

7 Conclusion

We have argued that the standard test section of the PDTB is too small to draw conclusions about whether a feature is generally useful or not, especially when using a larger label set, as is the case in recent work using second level labels. While these ideas are far from new and apply also to other NLP tasks with small evaluation sets, we think it is important to discuss this issue, as recent work in the field of discourse relation analysis has mostly ignored the issue of small test set sizes in the PDTB. Our experiments support our claim by showing that features that may look like they improve performance on the 11-way classification on the standard test set, did not always show a consistent improvement when the training / testing was split up differently. This means that we run a large risk of drawing incorrect conclusions about which features are helpful if we only stick out our small standard test set for evaluation.

8 Acknowledgements

This research was funded by the German Research Foundation (DFG) as part of SFB 1102 "Information Density and Linguistic Encoding". We also thank the anonymous reviewers for their careful reading and insightful comments.

[2]https://keras.io/

References

Yoshua Bengio and Yves Grandvalet. 2005. Bias in estimating the variance of k-fold cross-validation. In *Statistical modeling and analysis for complex data problems*, pages 75–95. Springer.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *journal of machine learning research*, volume3:1137–1155.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in nlp. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

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.

Chloé Braud and Pascal Denis. 2015. Comparing word representations for implicit discourse relation classification. In *Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 2201–2211, Lisbonne, Portugal. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Hugo Hernault, Danushka Bollegala, and Mitsuru Ishizuka. 2010. A semi-supervised approach to improve classification of infrequent discourse relations using feature vector extension. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 399–409, MIT, Massachusetts, USA. Association for Computational Linguistics.

Christopher Hidey and Kathleen McKeown. 2016. Identifying causal relation using parallel wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1424–1433, Berlin,Germany. Association for Computational Linguistics.

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*, pages 977–986, Baltimore, Maryland, USA. Association for Computational Linguistics.

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*, volume3:329–344.

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

Ji-Hyun Kim. 2009. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, volume53(11):3735–3745.

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

Thomas Meyer, Najeh Hajlaoui, and Andrei Popescu-Belis. 2015. Disambiguating discourse connectives for statistical machine translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(7):1184–1197.

Claudiu Mihăilă and Sophia Ananiadou. 2014. Semi-supervised learning of causal relations in biomedical scientific discourse. *Biomedical engineering online*, 13(2):1.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg 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 *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 108–112, Seoul, South Korea. Association for Computational Linguistics.

Emily Pitler, Mridhula Raghupathy, Hena Mehta, Ani Nenkova, Alan Lee, and Aravind K. Joshi. 2008. Easily identifiable discourse relations. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING2008)*, pages 85–88, Manchester, UK.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 683–691, Suntec, Singapore. 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 *In Proceedings of the 6th International Conference on Language Resources and Evaluation(LREC)*, Marrakech, Morocco. European Language Resources Association.

Chris Quirk and Hoifung Poon. 2016. Distant supervision for relation extraction beyond the sentence boundary. *arXiv preprint arXiv:1609.04873*.

Ines Rehbein, Merel Scholman, and Vera Demberg. 2016. Annotating discourse relations in spoken language: A comparison of the pdtb and ccr frameworks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 23–28, Portoro, Slovenia. European Language Resources Association.

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

Attapol Rutherford and Nianwen Xue. 2015. Improving the inference of implicit discourse relations via classifying explicit discourse connectives. In *Proceedings of the NAACL-HLT*, pages 799–808, Denver, Colorado. Association for Computational Linguistics.

Attapol Rutherford, Vera Demberg, and Nianwen Xue. 2017. A systematic study of neural discourse models for implicit discourse relation. In *European Chapter of the Association for Computational Linguistics.(EACL)*, Valencia, Spain. Association for Computational Linguistics.

Radu Soricut and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 149–156, Edmonton. Association for Computational Linguistics.

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

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad Christopher Bryant, and Attapol T. Rutherford. 2015. The conll-2015 shared task on shallow discourse parsing. In *Proceedings of Conference on Computational Natural Language Learning: Shared Task*, pages 1–16, Beijing, China. 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. pages 1–19, Berlin, Germany. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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, Lisbon, Portugal. Association for Computational Linguistics.

Biao Zhang, Deyi Xiong, and Jinsong Su. 2016. Neural discourse relation recognition with semantic memory. *arXiv preprint arXiv:1603.03873*.

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 *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1507–1514, Beijing, China. Association for Computational Linguistics.