Attention for Implicit Discourse Relation Recognition

Andre Cianflone, Leila Kosseim
Department of Engineering and Computer Science
Concordia University
{a_cianfl, kosseim}@encs.concordia.ca

Abstract

Implicit discourse relation recognition remains a challenging task as state-of-the-art approaches reach F1 scores ranging from 9.95% to 37.67% on the 2016 CoNLL shared task. In our work, we explore the use of a neural network which exploits the strong correlation between pairs of words across two discourse arguments that implicitly signal a discourse relation. We present a novel approach to Implicit Discourse Relation Recognition that uses an encoder-decoder model with attention. Our approach is based on the assumption that a discourse argument is "generated" from a previous argument and conditioned on a latent discourse relation, which we detect. Experiments show that our model achieves an F1 score of 38.25% on fine-grained classification, outperforming previous approaches and performing comparatively with state-of-the-art on coarse-grained classification, while computing alignment parameters without the need for additional pooling and fully connected layers.

Keywords: discourse relation recognition, sequence-to-sequence

1. Introduction

Shallow discourse relation recognition refers to the automatic identification of the relation between two segments of text. For example in:

(1) *I will go to Scotland* after **I complete my studies.**

The underlined discourse connective connects the first discourse argument (in italic) to the second discourse argument (in bold) via a *temporal* relation. Connectives constitute strong signals to identify discourse relations. In fact, given two arguments and a discourse connective many discourse parsers at the 2016 CoNLL Shared Task on Multilingual Shallow Discourse Parsing (SDP) (Xue et al., 2016) were around 78% accurate in recognizing the discourse relation on the SDP blind dataset. On the other hand, in implicit relations no connective is used. This is the case in:

(2) *I need to file my taxes.***Tomorrow is the deadline.**

In (2) the connective *because* is implied and the *contingency* relation is understood by the context. Unfortunately when the connective is absent, identifying the relation automatically becomes much more challenging. At the same 2016 CoNLL SDP shared task, the best implicit discourse relation (IDR) score on the blind test set without connectives reached 37.67% (Xue et al., 2016). In this paper we present a model to automatically recognize implicit discourse relations using an encoder-decoder with attention, a cross-argument word-pair alignment statistic in this context. We show that our model, with an F1 score of 38.25, outperforms other approaches on fine-grained classification, while performing comparatively with the state-of-the-art on coarse-grained classification.

2. Previous Work

Beginning with (Zhang et al., 2015a) and notably in the past year with the CoNLL SDP (Xue et al., 2016), neural network techniques have been used for IDR. Most of these models are based on convolutional neural networks (CNN), inspired by (Zhang et al., 2015a) and other work on sentence classification with CNN (such as (Kim, 2014; Zhang et al., 2015b)). The insight into these many works is that neural networks are better suited at capturing semantic clues between the two arguments of an implicit relation than traditional methods heavily reliant on feature engineering, as in (Pitler et al., 2009; Xue et al., 2015).

Given our correlation assumption, we sought a model that could successfully identify and exploit word pairs across arguments that are strong signals of a discourse relation, leading us to explore attention models. Although several neural network approaches have been proposed for IDR, to our knowledge none have investigated the use of encoder-decoder models with attention, an approach successfully applied to many applications including machine translation (Bahdanau et al., 2015), coreference resolution (Lee et al., 2017) and cloze-style reading comprehension (Cui et al., 2017). To improve translation, notably for longer sentences, a neural translation model is augmented with an attention mechanism uniquely purposed for capturing alignment (Bahdanau et al., 2015). The alignment model scores how well the input words from the source language match output words in the target language. Inspired by recent advances in the use of attention, we used attention to detect alignment scoring for IDR as word-pair features have be shown to contribute to IDR (Pitler et al., 2009; Biran and McKeown, 2013). However, unlike these methods we make no feature engineering. (Rönnqvist et al., 2017) also uses an attention mechanism to recognize implicit discourse relations. However, their approach differs from ours in two important ways: in (Rönnqvist et al., 2017), the two discourse arguments are concatenated to form a single input and the attention mechanism is applied over the entire input, which is fundamentally different to our sequence-to-sequence approach. Furthermore, their work is evaluated on the Chinese Discourse Treebank (Zhou and Xue, 2012).

Top Level	Nb Implicit Instances
Temporal	950
Contingency	4185
Comparison	2832
Expansion	8861
Total	**16828**

Table 1: Top-level breakdown of the PDTB with *entrel* merged into *expansion*

3. Datasets & Tasks

3.1. Datasets

Following the standard in the field, we used both the PDTB and the CoNLL SDP datasets. The PDTB dataset (Rashmi Prasad, 2008) contains 40,600 annotated discourse relations and their arguments over the 1 million word Wall Street Journal (WSJ) corpus (Prasad et al., 2008). The dataset includes four top-level classes of discourse relations; *temporal*, *contingency*, *comparison* and *expansion*; as well as level 2 and lever 3 types. For example, in the PDTB:

(3) *USAir has great promise.* **By the second half of 1990, USAir stock could hit 60.**

is labeled as "Contingency.Cause.Reason". A fifth top-level relation, *entrel* (short for *entity-based coherence*), is also defined but has no lower-level types. Table 1 shows statistics of the PDTB dataset.

The CoNLL SDP dataset consists of the full PDTB dataset with a minor reduction in the number of subtypes (Xue et al., 2016). Additionally, the SDP dataset includes a *blind test set*, a second test set created specifically for the 2015 and 2016 editions of the shared task. The blind test set consists of newswire text selected from English Wikinews[1] consistent with WSJ-style text and manually annotated with discourse relations and connectives (Xue et al., 2015).

3.2. Tasks

Given the difficulty of automatic IDR, most work focuses only on top-level classification; i.e. classifying only the four top-level relations with *entrel* merged into *expansion* as preferred by (Pitler et al., 2009; Rutherford and Xue, 2014; Ji and Eisenstein, 2015). The standard WSJ section breakdown is to use sections 2-20 for training, sections 21-22 for testing, and the other sections for development. Given the unbalanced dataset, as shown in Table 1, the task has traditionally been formulated as four binary classifiers. For the development and test sets, the negative samples consist of all other relations. The training set is evenly balanced between positive and negative where negatives samples are randomly drawn from WSJ sections 2 to 20 (excluding positives).

A notable exception to only top-level IDR was the 2015 and 2016 edition of the CoNLL SDP, which included fine-grained non-explicit discourse relation recognition.[2] The fine-grained task is to recognize the 16 low-level subtypes with a single classifier. Additionally, the WSJ section breakdown is different compared to the top-level dataset. The SDP training set consists of WSJ sections 2-21, section 22 for development, and section 23 for testing.

4. Our Model

We describe our model in two modules, the encoder-decoder Recurrent Neural Network (RNN) with attention and two varieties of the classifier.

4.1. Encoder-Decoder RNN with Attention

The standard encoder (Cho et al., 2014) encodes an input vector $\mathbf{x}$, where $\mathbf{x}$ is represented as a sequence of word embedding vectors, into a single context vector $c = q(h_1, \ldots, h_{T_x})$ and hidden state $h_t = f(x_t, h_{t-1})$. Functions f and q are nonlinearities, in our case Bidirectional RNN (Schuster and Paliwal, 1997) of type long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997).

Normally, the decoder predicts a sequence of words y_t where each y_t prediction is conditioned on past predictions and context vector c, maximizing the following joint probability:

$$p(y) = \prod_{t=1}^{T} p(y_t|\{y_1, \ldots, y_{t-1}\}, c) \tag{1}$$

In the context of RNNs, the conditional probability of each y_t in the joint probability of Eq.1 is modeled as a nonlinear function g with input y_t, context vector c and hidden state s_t:

$$p(y_t|\{y_1, \ldots, y_{t-1}\}, c) = g(y_{t-1}, s_t, c) \tag{2}$$

(Bahdanau et al., 2015) propose a unique context vector c_i for each decoding time step, redefining the decoder conditional probability for each word y_i as:

$$p(y_i|y_1, \ldots, y_{i-1}, \mathbf{x}) = g(y_{i-1}, s_i, c_i) \tag{3}$$

The context vector c_i is a weighted sum over all input hidden states $(h_1, \ldots, h_T)$:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \tag{4}$$

where weights $a_{ij} = \text{softmax}(e_{ij})$, $e_{ij} = a(s_{i-1}, h_j)$ and a is a feedforward neural network.

Using attention leads to a vectorized representation of the second argument (decoder output) which is not only informed of its context but also of its alignment, unlike Gated Relevance Networks (GRN) (Chen et al., 2016) where the arguments are not informed of the alignment. In the case of GRN, the two discourse arguments are vectorized with separate RNN layers (no interaction), followed by relevance layers (that compute word-pair interaction), and finally pooling and fully connected layers.

[1] https://en.wikinews.org

[2] Non-explicit discourse includes types *implicit*, *entrel*, and a third *altlex*, short for *alternative lexicalization*. Only a small fraction of the dataset, around 3%, consists of *altlex*. For this reason we will not discuss *altlex* and consider the terms "non-explicit" and "implicit" discourse interchangeably.

4.2. Classifiers

Given our classification task and since the decoder inputs (the second discourse argument words) are known and not predicted, the model is not trained by maximizing the likelihood of the decoder targets, as in Eq.1, but rather by minimizing the cross-entropy error between the predicted label $\hat{y}$ and the true label y for all possible labels l:

$$E(y, \hat{y}) = -\sum_{i=1}^{l} y_i \log(\hat{y}) \tag{5}$$

We experimented with two classifiers to predict $\hat{y}$. In the simplest case:

$$\hat{y} = f(Wh_T + b) \tag{6}$$

where f is the softmax function, h_T is the final decoder hidden state of size d, $W \in \mathbb{R}^{l \times d}$ is a parameter matrix and $b \in \mathbb{R}^l$ a bias vector. In this case the classifier only relies on the last hidden state, minimizing the total number of parameters at the expense of information loss. We denote this Classifier with Attention CA, shown in Figure 1.

In the second classifier, $\hat{y}$ is a function of:

$$p = \max_{t=1}^{T}(h_{dec_1}, \ldots, h_{dec_t}) \tag{7}$$

$$h = g(W_d p + b_d) \tag{8}$$

$$\hat{y} = f(W_s h + b_s) \tag{9}$$

where p is a T sized concatenated vector of the maximum values over each decoder hidden state h_{dec}, i.e. 1D max pooling. $W_d \in \mathbb{R}^{v \times T}$ and $Ws \in \mathbb{R}^{l \times v}$ are parameter matrices, $b \in \mathbb{R}^v$ and $b \in \mathbb{R}^l$ are bias terms, and g a non-linearity. In this case each decoded time step informs the relation classification. We denote this Classifier from Sequence with Attention CSA, as shown in Figure 2

5. Experiments

In this section we outline our data preprocessing and experiments. The raw texts from the PDTB and the CoNLL SDP are converted to lower case and tokenized. Then we keep only the 10,000 most common words. After forming a dictionary of unique tokens, we substitute each token with a dense word embedding from a pretrained model. Following the preferred embeddings used at the 2016 CoNLL SDP (Xue et al., 2016), we used the 300 dimensional pre-trained Word2Vec binaries[3], trained by continuous skip-gram (Mikolov et al., 2013) for both top-level and fine-grained classification. While the PDTB samples contain additional data such as part-of-speech tags and parse trees, no additional data is used.

The top-level classification consists of four separately trained binary classifiers, while we train a single classifier for the fine-grained classification. We experiment using LSTM and GRU (Cho et al., 2014) cells, opting for LSTM since it showed slightly better results. The number of cell parameters were randomly searched at each training run. We randomly switched between bidirectional encoder or single direction. For the CSA, we additionally performed hyper-parameter search on the number of hidden units. Our main parameters that produced the best performance are listed in Table 2. Our models were optimized with the Adam algorithm (Kingma and Ba, 2015). Models evaluated on the test sets are based on optimal validation set F1 score.

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

Model		Parameter	Value
CSA	CA	batch size	32
		embedding size	300
		cell type	LSTM
		cell units	100
		pooling	1D max
		dense layer units	60

Table 2: Architecture parameters. Dense layer refers to the CSA model's fully connected layer between pooling and softmax layers.

ID	Author	Blind	Test	Dev
ecnucs	Wang	34.18	40.91	46.40
tbmihaylov	Mihaylov	34.51	39.19	40.32
tao0920	Qin	35.38	38.20	46.33
gtnlp	n/a	36.75	34.95	40.72
ttr	Rutherford	37.67	36.13	40.32
CSA	ours	35.07	28.05	36.58
CA	ours	**38.25**	35.63	39.42

Table 3: F1 scores of fine-grained IDR compared to top 5 teams. (Wang and Lan, 2016; Mihaylov and Frank, 2016; Qin et al., 2016; Rutherford and Xue, 2016)

6. Results & Analysis

Given the unbalanced datasets, performance is evaluated solely on F1 scores. Table 3, summarizes our top-level classification results on the PDTB dataset in comparison with other authors and Table 4 our fine-grained classification results[4] on the CoNLL SDP dataset.

As shown in Table 3, our CA model scored 38.25% on the fine-grained classification, over state-of-the-art F1 score of 37.67%. Observing the blind test set results in Table 3 we note how our model generalizes well to a different dataset (Wikinews). Other top models such as "gtnlp" and "ecnucs" have a more than 10 point difference between the development score and blind test score compared to 2 points in the CA case.

For the top-level classification, our CA model (see Table 4) scored well in the case of *expansion* with 80.72% F1 score, the largest relation class, and *contingency*, while*temporal* was better than most other approaches. The F1 of 30.56% for *comparison* was far from the top result in Table 4, likely due to the small dataset size.

It is interesting to note that the results achieved by the CA model are based on a relatively shallow, single bidirectional RNN encoder layer and single RNN decoder layer with attention. It is possible that the chosen input embedding had a minor impact on our results. We would have liked to measure the embedding effect to compare with (Chen et al., 2016), but to our knowledge the embedding is not publicly available.

[4]We used the official CoNLL scorer for comparison: https://github.com/attapol/conll16st

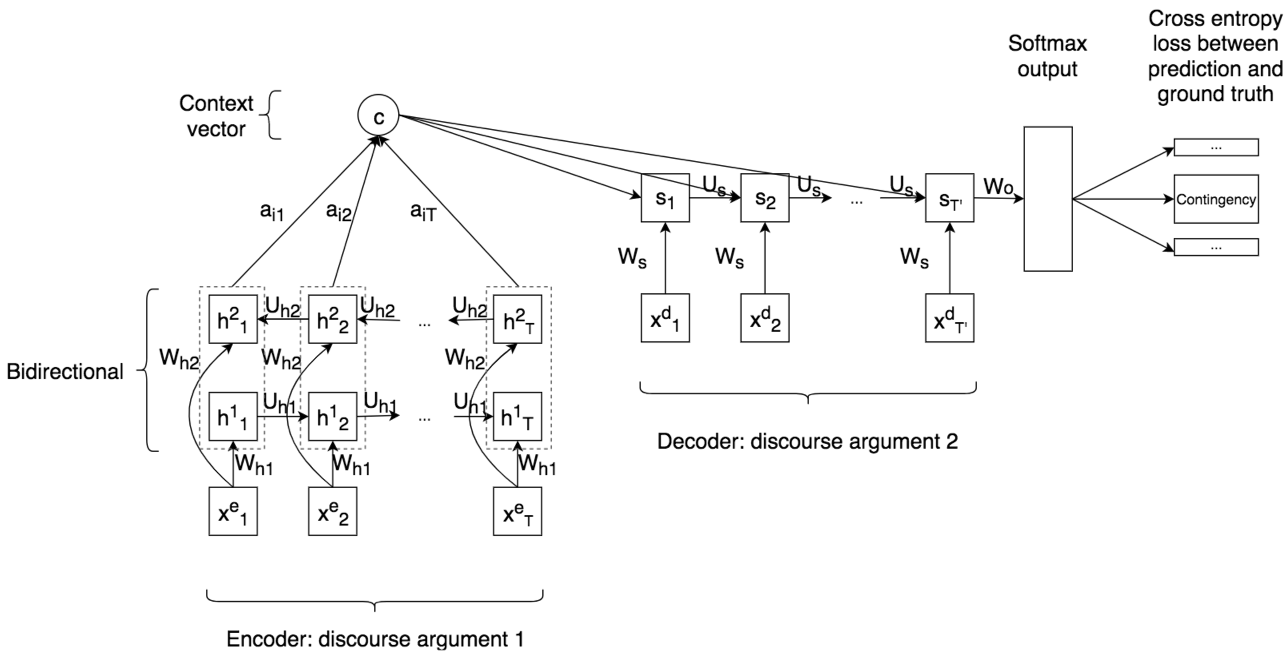

Figure 1: Our classifier with attention (CA): an encoder-decoder recurrent neural network with attention with the last hidden state used for classification. In the doted rectangles, the forward and backward hidden states are concatenated. Note there is no backpropagation through time from output predictions at each time step. Only the final cross-entropy error is backpropagated through time.

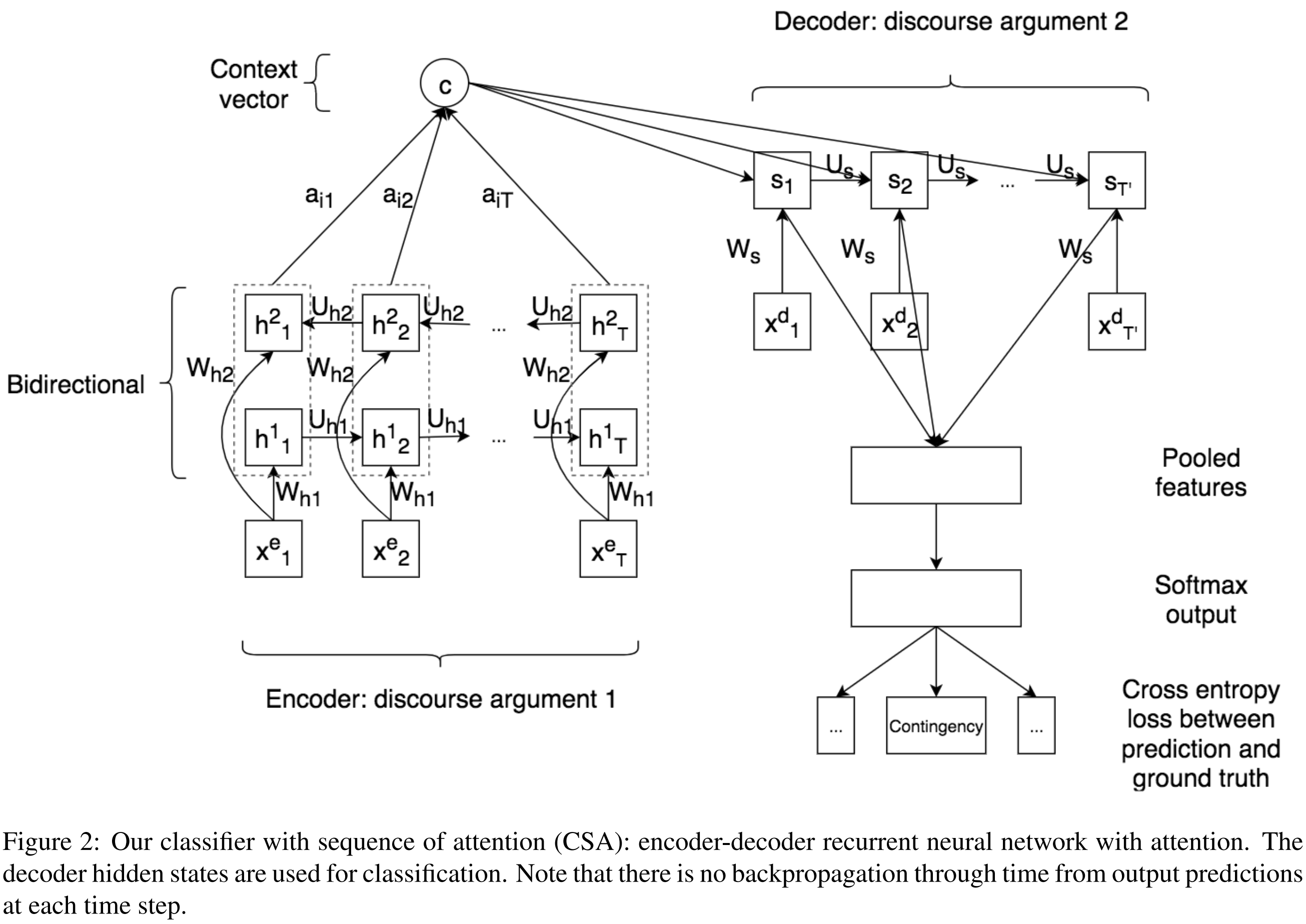

Figure 2: Our classifier with sequence of attention (CSA): encoder-decoder recurrent neural network with attention. The decoder hidden states are used for classification. Note that there is no backpropagation through time from output predictions at each time step.

Author	Comp.	Cont.	Exp.	Temp.
Pitler	21.96	47.13	76.42	16.76
Zhou	31.79	47.16	70.11	20.30
Park	31.32	49.82	79.22	26.57
Rutherford	39.70	54.42	80.44	28.69
Ji	35.93	52.78	80.02	27.63
Chen	**40.17**	54.76	80.62	**31.32**
CSA	27.02	49.86	77.45	24.43
CA	30.56	**54.80**	**80.72**	27.15

Table 4: F1 scores of top-level IDR for: *comparison, contingency, expansion, temporal*. Note that *entrel* is merged into *expansion*, as done in previous works. (Pitler et al., 2009; Zhou et al., 2010; Park and Cardie, 2012; Rutherford and Xue, 2014; Ji and Eisenstein, 2015; Chen et al., 2016)

We were surprised by the CSA's lower performance in all cases. We believed the model would be more robust if the classification layer had inputs from all decoded hidden states directly. However, using only the final state vector resulted in higher classification score while using less parameters. This may be due to overfitting. We would need to reevaluate the model on a larger dataset.

7. Conclusion

We presented an efficient encoder-decoder model with attention for implicit discourse relation recognition. Our model computes attention between discourse argument word pairs without feature engineering and without the need for additional fully connected layers, minimizing the number of trainable parameters. Finally, we show that our model generalizes well to unseen datasets on fine-grained classification, outperforming state-of-the-art without large variance in scoring between development and test sets, and outperforms in two categories in the coarse-grained case. In future work we would like to explore in more detail automatically learned alignment for IDR and text generation based on these models.

Acknowledgement

The authors would like to thank the anonymous reviewers for their feedback on the paper. We gratefully acknowledge the financial support of NSERC and the NVIDIA Corporation for the donation of the Titan X Pascal GPU used for this research.

8. Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Advances in Neural Information Processing Systems (NIPS 2015)*, pages 649–657, Montreal, Canada, December.

Biran, O. and McKeown, K. (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.

Chen, J., Zhang, Q., Liu, P., Qiu, X., and Huang, X. (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-2016)*, pages 1726–1735.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. pages 1724â–1734, Doha, Qatar.

Cui, Y., Chen, Z., Wei, S., Wang, S., Liu, T., and Hu, G. (2017). Attention-over-attention neural networks for reading comprehension. pages 593–602.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Ji, Y. and Eisenstein, J. (2015). One vector is not enough: Entity-augmented distributional semantics for discourse relations. *Transactions of the Association for Computational Linguistics*, pages 329–344, June.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceeding of the 2015 International Conference on Learning Representation (ICLR 2015)*, San Diego, California.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution.

Mihaylov, T. and Frank, A. (2016). Discourse relation sense classification using cross-argument semantic similarity based on word embeddings. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task (CoNLL)*, pages 100–107, Berlin, Germany.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems (NIPS 2013)*, pages 3111–3119, Lake Tahoe, USA, December.

Park, J. and Cardie, C. (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. Association for Computational Linguistics.

Pitler, E., Louis, A., and Nenkova, A. (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 Processinf (ACL/IJCNLP). Volume 2*, pages 683–691. Association for Computational Linguistics.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. K., and Webber, B. L. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Qin, L., Zhang, Z., and Zhao, H. (2016). Shallow discourse parsing using convolutional neural network. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task (CoNLL)*, pages 70–77, Berlin, Germany.

Rönnqvist, S., Schenk, N., and Chiarcos, C. (2017). A recurrent neural model with attention for the recognition of chinese implicit discourse relations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 256–262.

Rutherford, A. and Xue, N. (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. Association for Computational Linguistics.

Rutherford, A. and Xue, N. (2016). Robust non-explicit neural discourse parser in english and chinese. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task (CoNLL)*, pages 55–59, Berlin, Germany.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Wang, J. and Lan, M. (2016). Two end-to-end shallow discourse parsers for english and chinese in conll-2016 shared task. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task (CoNLL)*, pages 33–40, Berlin, Germany.

Xue, N., Ng, H. T., Pradhan, S., Bryant, R. P. C., and Rutherford, A. T. (2015). The conll-2015 shared task on shallow discourse parsing. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task (CoNLL)*, pages 1–16, Beijing, China.

Xue, N., Ng, H. T., Rutherford, A., Webber, B., Wang, C., and Wang, H. (2016). Conll 2016 shared task on multilingual shallow discourse parsing. *Proceedings of the Twentieth Conference on Computational Natural Language Learning: Shared Task (CoNLL)*, pages 1–19.

Zhang, B., Su, J., Xiong, D., Lu, Y., Duan, H., and Yao, J. (2015a). 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, September. Association for Computational Linguistics.

Zhang, X., Zhao, J., and LeCun, Y. (2015b). Character-level convolutional networks for text classification. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS 2015)*, pages 649–657, Montreal, Canada, December.

Zhou, Y. and Xue, N. (2012). Pdtb-style discourse annotation of chinese text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: (ACL)*, pages 69–77. Association for Computational Linguistics.

Zhou, Z.-M., Xu, Y., Niu, Z.-Y., Lan, M., Su, J., and Tan, C. L. (2010). Predicting discourse connectives for implicit discourse relation recognition. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 1507–1514. Association for Computational Linguistics.

Language Resource References

Rashmi Prasad, Alan Lee, Nikhil Dinesh, Eleni Miltsakaki, Geraud Campion, Aravind Joshi, Bonnie Webber. (2008). *Penn Discourse Treebank Version 2.0*. Linguistic Data Consortium, 2.0, ISLRN 488-589-036-315-2.