A Recurrent Neural Model with Attention for the Recognition of Chinese Implicit Discourse Relations

Samuel Rönnqvist[1,2,*], **Niko Schenk**[2,*] and **Christian Chiarcos**[2]
[1]Turku Centre for Computer Science – TUCS, Åbo Akademi University, Turku, Finland
[2]Applied Computational Linguistics Lab, Goethe University, Frankfurt am Main, Germany
sronnqvi@abo.fi
{schenk,chiarcos}@informatik.uni-frankfurt.de

Abstract

We introduce an attention-based Bi-LSTM for Chinese implicit discourse relations and demonstrate that modeling argument pairs as a joint sequence can outperform word order-agnostic approaches. Our model benefits from a partial sampling scheme and is conceptually simple, yet achieves state-of-the-art performance on the Chinese Discourse Treebank. We also visualize its attention activity to illustrate the model's ability to selectively focus on the relevant parts of an input sequence.

1 Introduction

True text understanding is one of the key goals in Natural Language Processing and requires capabilities beyond the lexical semantics of individual words or phrases. Natural language descriptions are typically driven by an inter-sentential coherent structure, exhibiting specific *discourse* properties, which in turn contribute significantly to the global meaning of a text. Automatically detecting how meaning units are organized benefits practical downstream applications, such as question answering (Sun and Chai, 2007), recognizing textual entailment (Hickl, 2008), sentiment analysis (Trivedi and Eisenstein, 2013), or text summarization (Hirao et al., 2013).

Various formalisms in terms of semantic coherence frameworks have been proposed to account for these contextual assumptions (Mann and Thompson, 1988; Lascarides and Asher, 1993; Webber, 2004). The annotation schemata of the Penn Discourse Treebank (Prasad et al., 2008, PDTB) and the Chinese Discourse Treebank (Zhou and Xue, 2012, CDTB), for instance, define discourse units as syntactically motivated character spans in the text, augmented with relations pointing from the second argument (*Arg2*, prototypically, a discourse unit associated with an explicit discourse marker) to its antecedent, i.e., the discourse unit *Arg1*. Relations are labeled with a relation type (its sense) and the associated discourse marker. Both, PDTB and CDTB, distinguish *explicit* from *implicit* relations depending on the presence of such a marker (e.g., *because*/ 因).[1] Sense classification for implicit relations is by far more challenging because the argument pairs lack the marker as an important feature. Consider, for instance, the following example from the CDTB as implicit CONJUNCTION:

Arg1: 会谈就一些原则和具体问题进行了深入讨论，达成了一些谅解 *In the talks, they discussed some principles and specific questions in depth, and reached some understandings*

Arg2: 双方一致认为会谈具有积极成果 *Both sides agree that the talks have positive results*

Motivation: Previous work on implicit sense labeling is heavily feature-rich and requires domain-specific, semantic lexicons (Pitler et al., 2009; Feng and Hirst, 2012; Huang and Chen, 2011). Only recently, resource-lean architectures have been proposed. These promising neural methods attempt to infer latent representations appropriate for implicit relation classification (Zhang et al., 2015; Ji et al., 2016; Chen et al., 2016). So far, unfortunately, these models have been evaluated *only* on four top-level senses—sometimes even with inconsistent evaluation setups.[2] Furthermore, most systems have initially been designed for the English PDTB and involve complex, task-

[*]Both first authors contributed equally to this work.

[1]The set of relation types and senses is completed by alternative lexicalizations (ALTLEX/discourse marker rephrased), and entity relations (ENTREL/anaphoric coherence).

[2]E.g., four binary classifiers vs. four-way classification.

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 256–262
Vancouver, Canada, July 30 - August 4, 2017.
https://doi.org/10.18653/v1/P17-2040

specific architectures (Liu and Li, 2016), while discourse modeling techniques for Chinese have received very little attention in the literature and are still seriously underrepresented in terms of publicly available systems. What is more, over 80% of all words in Chinese discourse relations are implicit—compared to only 52% in English (Zhou and Xue, 2012).

Recently, in the context of the CoNLL 2016 shared task (Xue et al., 2016), a first independent evaluation platform beyond class level has been established. Surprisingly, the best performing neural architectures to date are standard *feedforward* networks, cf. Wang and Lan (2016); Schenk et al. (2016); Qin et al. (2016). Even though these specific models completely ignore word order within arguments, such feedforward architectures have been claimed by Rutherford et al. (2016) to generally outperform any thoroughly-tuned recurrent architecture.

Our Contribution: In this work, we release the first attention-based *recurrent* neural sense classifier, specifically developed for Chinese implicit discourse relations. Inspired by Zhou et al. (2016), our system is a practical adaptation of the recent advances in relation modeling extended by a novel sampling scheme.

Contrary to previous assertions by Rutherford et al. (2016), our model demonstrates superior performance over traditional bag-of-words approaches with feedfoward networks by treating discourse arguments as a joint sequence. We evaluate our method within an independent framework and show that it performs very well beyond standard class-level predictions, achieving state-of-the-art accuracy on the CDTB test set.

We illustrate how our model's attention mechanism provides means to highlight those parts of an input sequence that are relevant for the classification decision, and thus, it may enable a better understanding of the implicit discourse parsing problem. Our proposed network architecture is flexible and largely language-independent as it operates only on word embeddings. It stands out due to its structural simplicity and builds a solid ground for further development towards other textual domains.

2 Approach

We propose the use of an attention-based bidirectional Long Short-Term Memory (Hochreiter and Schmidhuber, 1997, LSTM) network to predict senses of discourse relations. The model draws upon previous work on LSTM, in particular its bidirectional mode of operation (Graves and Schmidhuber, 2005), attention mechanisms for recurrent models (Bahdanau et al., 2014; Hermann et al., 2015), and the combined use of these techniques for entity relation recognition in annotated sequences (Zhou et al., 2016). More specifically, our model is a flexible recurrent neural network with capabilities to *sequentially* inspect tokens and to highlight which parts of the input sequence are most informative for the discourse relation recognition task, using the weighting provided by the attention mechanism. Furthermore, the model benefits from a novel sampling scheme for arguments, as elaborated below. The system is learned in an end-to-end manner and consists of multiple layers, which are illustrated in Figure 1.

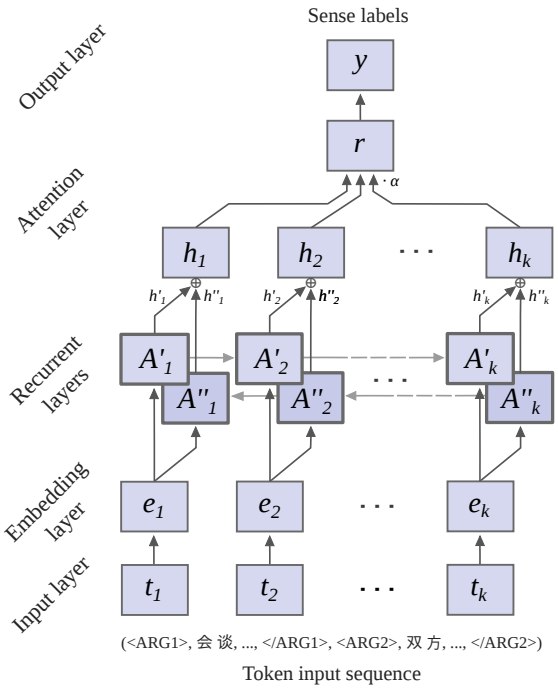

Figure 1: The attention-based bidirectional LSTM network for the task of modeling argument pairs for Chinese implicit discourse relations.

First, token sequences are taken as input and special markers (<ARG1>, </ARG1>, etc.) are inserted into the corresponding positions to inform the model on the start and end points of argument spans. This way, we can ensure a general flexibility in modeling discourse units and could easily extend them with additional context, for instance. In our experiments on implicit arguments,

only the tokens in the respective spans are considered. Note that, unlike previous works, our approach models *Arg1-Arg2* pairs as a *joint* sequence and does not first compute intermediate representations of arguments separately.

Second, an input layer encodes tokens using one-hot vector representations (t_i for tokens at positions $i \in [1, k]$), and a subsequent embedding layer provides a dense representation (e_i) to serve as input for the recurrent layers. The embedding layer is initialized using pre-trained word vectors, in our case 300-dimensional Chinese Gigaword vectors (Graff and Chen, 2005).[3] These embeddings are further tuned as the network is trained towards the prediction task. Embeddings for unknown tokens, e.g., markers, are trained by back-propagation only. Note that, tokens, markers and the pre-trained vectors represent the only source of information for the prediction task.

For the recurrent setup, we use a layer of LSTM networks in a bidirectional manner, in order to better capture dependencies between parts of the input sequence by inspection of both left and right-hand-side contexts at each time step. The LSTM holds a state representation as a continuous vector passed to the subsequent time step, and it is capable of modeling long-range dependencies due to its gated memory. The forward (A') and backward (A'') LSTMs traverse the sequence e_i, producing sequences of vectors h'_i and h''_i respectively, which are then summed together (indicated by $\oplus$ in Figure 1).

The resulting sequence of vectors h_i is reduced into a single vector and fed to the final softmax output layer in order to classify the sense label y of the discourse relation. This vector may be obtained either as the final vector h produced by an LSTM, or through pooling of all h_i, or by using attention, i.e., as a weighted sum over h_i. While the model may be somewhat more difficult to optimize using attention, it provides the added benefit of interpretability, as the weights highlight to what extent the classifier considers the LSTM state vectors at each token during modeling. This is particularly interesting for discourse parsing, as most previous approaches have provided little support for pinpointing the driving features in each argument span.

Finally, the attention layer contains the trainable vector w (of the same dimensionality as vectors h_i) which is used to dynamically produce a weight vector α over time steps i by:

$$\alpha = softmax(w^T tanh(H))$$

where H is a matrix consisting of vectors h_i. The output layer r is the weighted sum of vectors in H:

$$r = H\alpha^T$$

Partial Argument Sampling: For the purpose of enlarging the instance space of training items in the CDTB, and thus, in order to improve the predictive performance of the model, we propose a novel *partial sampling* scheme of arguments, whereby the model is trained and validated on sequences containing both arguments, as well as *single* arguments. A data point (a_1, a_2, y), with a_i being the token sequence of argument i, is expanded into $\{(a_1, a_2, y), (a_1, a_2, y), (a_1, y), (a_2, y)\}$. We duplicate bi-argument samples (a_1, a_2, y) (in training and development data only) to balance their frequencies against single-argument samples.

Two lines of motivation support the inclusion of single argument training examples, grounded in linguistics and machine learning, respectively. First, it has been shown that single arguments in isolation can evoke a strong expectation towards a certain implicit discourse relation, cf. Asr and Demberg (2015) and, in particular, Rohde and Horton (2010) in their psycholinguistic study on *implicit causality verbs*. Second, the procedure may encourage the model to learn better representations of individual argument spans in support of modeling of arguments in composition, cf. LeCun et al. (2015). Due to these aspects, we believe this data augmentation technique to be effective in reinforcing the overall robustness of our model.

Implementational Details: We train the model using fixed-length sequences of 256 tokens with zero padding at the beginning of shorter sequences and truncate longer ones. Each LSTM has a vector dimensionality of 300, matching the embedding size. The model is regularized by 0.5 dropout rate between the layers and weight decay ($2.5e^{-6}$) on the LSTM inputs. We employ Adam optimization (Kingma and Ba, 2014) using the cross-entropy loss function with mini batch size of 80.[4]

[3] http://www.cs.brandeis.edu/~clp/conll16st/dataset.html

[4] The model is implemented in *Keras* https://keras.io/.

CDTB Development Set			CDTB Test Set		
Rank	System	% accuracy	Rank	System	% accuracy
1	Wang and Lan (2016)	73.53	1	Wang and Lan (2016)	72.42
2	Qin et al. (2016)	71.57	2	Schenk et al. (2016)	71.87
3	Schenk et al. (2016)	70.59	3	Rutherford and Xue (2016)	70.47
4	Rutherford and Xue (2016)	68.30	4	Qin et al. (2016)	67.41
5	Weiss and Bajec (2016)	66.67	5	Weiss and Bajec (2016)	64.07
6	Weiss and Bajec (2016)	61.44	6	Weiss and Bajec (2016)	63.51
7	Jian et al. (2016)	21.90	7	Jian et al. (2016)	21.73
	This Paper:	**93.52***		**This Paper:**	**73.01**

Table 1: Non-explicit parser scores on the official CoNLL 2016 CDTB development and test sets. (*Scores on development set are obtained through partial sampling and are not directly comparable.)

Sense Label	Training	Dev't	Test
CONJUNCTION	5,174	189	228
majority class	(66.3%)	(62.8%)	(64.8%)
EXPANSION	1,188	48	40
ENTREL	1,099	50	71
CAUSATION	187	10	8
CONTRAST	66	3	1
PURPOSE	56	1	3
CONDITIONAL	26	0	1
TEMPORAL	26	0	0
PROGRESSION	7	0	0
# impl. rels	7,804	301	352

Table 2: Implicit sense labels in the CDTB.

3 Evaluation

We evaluate our recurrent model on the CoNLL 2016 shared task data[5] which include the official training, development and test sets of the CDTB; cf. Table 2 for an overview of the implicit sense distribution.[6]

In accordance with previous setups (Rutherford et al., 2016), we treat entity relations (ENTREL) as implicit and exclude ALTLEX relations. In the evaluation, we focus on the *sense-only* track, the subtask for which gold arguments are provided and a system is supposed to label a given argument pair with the correct sense. The results are shown in Table 1.

With our proposed architecture it is possible to correctly label 257/352 (73.01%) of implicit relations on the test set, outperforming the best feed-forward system of Wang and Lan (2016) and all other word order-agnostic approaches. Development and test set performances suggest the robustness of our approach and its ability to generalize to unseen data.

Ablation Study: We perform an ablation study to quantitatively assess the contribution of two of the characteristic aspects of our model. First, we compare the use of the attention mechanism against the simpler alternative of feeding the final LSTM hidden vectors (h'_k and h''_1) directly to the output layer. When attention is turned off, this yields an absolute decrease in performance of 2.70% on the test set, which is substantial and significant according to a Welch two-sample t-test ($p < .001$). Second, we independently compare the use of the partial sampling scheme against training on the standard argument pairs in the CDTB. Here, the absence of the partial sampling scheme yields an absolute decrease in accuracy of 5.74% ($p < .001$), which demonstrates its importance for achieving competitive performance on the task.

Performance on the PDTB: As a side experiment, we investigate the model's language independence by applying it to the implicit argument pairs of the English PDTB. Due to computational time constraints we do not optimize hyperparameters, but instead train the model using identical settings as for Chinese, which is expected to lead to suboptimal performance on the evaluation data. Nevertheless, we measure 27.09% accuracy on the PDTB test set (surpassing the majority class baseline of 22.01%), which shows that the model has potential to generalize across implicit discourse relations in a different language.

[5] http://www.cs.brandeis.edu/~clp/conll16st/

[6] Note that, in the CDTB, implicit relations appear almost *three times more often* than explicit relations. Out of these, 65% appear within the same sentence. Finally, 25 relations in the training set have two labels.

Figure 2: Visualization of attention weights for Chinese characters with high (dark blue) and low (light blue) intensities. The underlined English phrases are semantically structure-shared by the two arguments.

Visualizing Attention Weights: Finally, in Figure 2, we illustrate the learned attention weights which pinpoint important subcomponents within a given implicit discourse relation. For the implicit CONJUNCTION relation the weights indicate a peak on the transition between the argument boundary, establishing a connection between the semantically related terms *understandings–agree*. Most ENTRELs show an opposite trend: here second arguments exhibit larger intensities than *Arg1*, as most entity relations follow the characteristic writing style of newspapers by adding additional information by reference to the same entity.

4 Summary & Outlook

In this work, we have presented the first attention-based recurrent neural sense labeler specifically developed for Chinese implicit discourse relations. Its ability to model discourse units sequentially and jointly has been shown to be highly beneficial, both in terms of state-of-the-art performance on the CDTB (outperforming word order-agnostic feedforward approaches), and also in terms of insightful observations into the inner workings of the model through its attention mechanism. The architecture is structurally simple, benefits from partial argument sampling, and can be easily adapted to similar relation recognition tasks. In future work, we intend to extend our approach to different languages and domains, e.g., to the recent data sets on narrative story understanding or question answering (Mostafazadeh et al., 2016; Feng et al., 2015). We believe that recurrent modeling of implicit discourse information can be a driving force in successfully handling such complex semantic processing tasks.[7]

Acknowledgments

The authors would like to thank Ayah Zirikly, Philip Schulz and Wei Ding for their very helpful suggestions on an early draft version of the paper, and also thank the anonymous reviewers for their valuable feedback and insightful comments. We are grateful to Farrokh Mehryary for technical support with the attention layer implementation. Computational resources were provided by CSC – IT Centre for Science, Finland, and Arcada University of Applied Sciences, Helsinki, Finland. Our research at Goethe University Frankfurt was supported by the project 'Linked Open Dictionaries (LiODi, 2015-2020)', funded by the German Ministry for Education and Research (BMBF).

[7]The code involved in this study is publicly available at http://www.acoli.informatik.uni-frankfurt.de/resources/.

References

Fatemeh Torabi Asr and Vera Demberg. 2015. Uniform Information Density at the Level of Discourse Relations: Negation Markers and Discourse Connective Omission. In *11th International Conference on Computational Semantics (IWCS)*. page 118. http://www.coli.uni-saarland.de/ fatemeh/iwcs2015.pdf.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473. http://arxiv.org/abs/1409.0473.

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 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. http://aclweb.org/anthology/P/P16/P16-1163.pdf.

Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015, Scottsdale, AZ, USA, December 13-17, 2015*. pages 813–820. https://doi.org/10.1109/ASRU.2015.7404872.

Vanessa Wei Feng and Graeme Hirst. 2012. Text-level Discourse Parsing with Rich Linguistic Features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '12, pages 60–68. http://www.aclweb.org/anthology/P12-1007.

David Graff and Ke Chen. 2005. Chinese Gigaword. LDC Catalog No.: LDC2003T09, ISBN, 1:58563–58230.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18(5-6):602–610. https://doi.org/10.1016/j.neunet.2005.06.042.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*. pages 1693–1701.

Andrew Hickl. 2008. Using Discourse Commitments to Recognize Textual Entailment. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '08, pages 337–344. http://dl.acm.org/citation.cfm?id=1599081.1599124.

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. 2013. Single-Document Summarization as a Tree Knapsack Problem. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*. pages 1515–1520. http://aclweb.org/anthology/D/D13/D13-1158.pdf.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Hen-Hsen Huang and Hsin-Hsi Chen. 2011. Chinese Discourse Relation Recognition. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, Chiang Mai, Thailand, pages 1442–1446. http://www.aclweb.org/anthology/I11-1170.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A Latent Variable Recurrent Neural Network for Discourse-Driven Language Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 332–342. http://www.aclweb.org/anthology/N16-1037.

Ping Jian, Xiaohan She, Chenwei Zhang, Pengcheng Zhang, and Jian Feng. 2016. Discourse Relation Sense Classification Systems for CoNLL-2016 Shared Task. In *Proceedings of the CoNLL-16 shared task*. Association for Computational Linguistics, pages 158–163. https://doi.org/10.18653/v1/K16-2022.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980. http://arxiv.org/abs/1412.6980.

Alex Lascarides and Nicholas Asher. 1993. Temporal Interpretation, Discourse Relations and Commonsense entailment. *Linguistics and Philosophy* 16(5):437–493.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521(7553):436–444.

Yang Liu and Sujian Li. 2016. Recognizing Implicit Discourse Relations via Repeated Reading: Neural Networks with Multi-Level Attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*. pages 1224–1233. http://aclweb.org/anthology/D/D16/D16-1130.pdf.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text* 8(3):243–281.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 839–849. http://www.aclweb.org/anthology/N16-1098.

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*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL 2009, pages 683–691. http://www.aclweb.org/anthology/P/P09/P09-1077.pdf.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings, 6th International Conference on Language Resources and Evaluation*. Marrakech, Morocco, pages 2961–2968.

Lianhui Qin, Zhisong Zhang, and Hai Zhao. 2016. Shallow Discourse Parsing Using Convolutional Neural Network. In *Proceedings of the CoNLL-16 shared task*. Association for Computational Linguistics, pages 70–77. https://doi.org/10.18653/v1/K16-2010.

Hannah Rohde and William Horton. 2010. Why or what next? Eye movements reveal expectations about discourse direction. Talk at the 23rd Annual CUNY Conference on Human Sentence Processing. New York, NY.

Attapol Rutherford and Nianwen Xue. 2016. Robust Non-Explicit Neural Discourse Parser in English and Chinese. In *Proceedings of the CoNLL-16 shared task*. Association for Computational Linguistics, pages 55–59. https://doi.org/10.18653/v1/K16-2007.

Attapol T. Rutherford, Vera Demberg, and Nianwen Xue. 2016. Neural Network Models for Implicit Discourse Relation Classification in English and Chinese without Surface Features. *CoRR* abs/1606.01990. http://arxiv.org/abs/1606.01990.

Niko Schenk, Christian Chiarcos, Kathrin Donandt, Samuel Rönnqvist, Evgeny Stepanov, and Giuseppe Riccardi. 2016. Do We Really Need All Those Rich Linguistic Features? A Neural Network-Based Approach to Implicit Sense Labeling. In *Proceedings of the CoNLL-16 shared task*. Association for Computational Linguistics, pages 41–49. https://doi.org/10.18653/v1/K16-2005.

Mingyu Sun and Joyce Y Chai. 2007. Discourse processing for context question answering based on linguistic knowledge. *Knowledge-Based Systems* 20(6):511–526.

Rakshit S. Trivedi and Jacob Eisenstein. 2013. Discourse connectors for latent subjectivity in sentiment analysis. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*. pages 808–813. http://aclweb.org/anthology/N/N13/N13-1100.pdf.

Jianxiang Wang and Man Lan. 2016. Two End-to-end Shallow Discourse Parsers for English and Chinese in CoNLL-2016 Shared Task. In *Proceedings of the CoNLL-16 shared task*. Association for Computational Linguistics, pages 33–40. https://doi.org/10.18653/v1/K16-2004.

Bonnie L. Webber. 2004. D-LTAG: extending lexicalized TAG to discourse. *Cognitive Science* 28(5):751–779. http://dblp.uni-trier.de/db/journals/cogsci/cogsci28.html.

Gregor Weiss and Marko Bajec. 2016. Discourse Sense Classification from Scratch using Focused RNNs. In *Proceedings of the CoNLL-16 shared task*. Association for Computational Linguistics, pages 50–54. https://doi.org/10.18653/v1/K16-2006.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Bonnie Webber, Attapol Rutherford, Chuan Wang, and Hongmin Wang. 2016. The CoNLL-2016 Shared Task on Shallow Discourse Parsing. In *Proceedings of the Twentieth Conference on Computational Natural Language Learning - Shared Task*. Association for Computational Linguistics, Berlin, Germany.

Biao Zhang, Jinsong Su, Deyi Xiong, Yaojie Lu, Hong Duan, and Junfeng Yao. 2015. Shallow Convolutional Neural Network for Implicit Discourse Relation Recognition. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*. pages 2230–2235. http://aclweb.org/anthology/D/D15/D15-1266.pdf.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. http://aclweb.org/anthology/P/P16/P16-2034.pdf.

Yuping Zhou and Nianwen Xue. 2012. PDTB-style Discourse Annotation of Chinese Text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Jeju Island, Korea, pages 69–77. http://www.aclweb.org/anthology-new/P/P12/P12-1008.bib.